1980

SPEAKER VERIFICATION FROM NOISY SPEECH.

MARIAN RICHARD. BARANIECKI
University of Windsor

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

SPEAKER VERIFICATION FROM NOISY SPEECH

by

Marian Richard Baranіecki

A Dissertation

Submitted to the Faculty of Graduate Studies
through the Department of Electrical Engineering
in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy at
the University of Windsor
Windsor, Ontario, Canada

1980

ABSTRACT

In this work the development of a speaker verification system based on an orthogonal linear prediction model, that would operate under different ambient conditions, is investigated.

The effectiveness of the orthogonal parameter model for speech utterances obtained under varying ambient conditions is evaluated.

The optimal orders of linear prediction are determined for various kinds of speech:

i) high quality speech

ii) the speech with additive wideband noise

iii) the telephone speech

The expression is obtained to estimate the orthogonal parameters of a clean signal in terms of parameters derived from noisy signal.

Methods for improving the accuracy of verification are developed and tested. The results obtained indicate that through the use of adaptive noise cancelling filter, high accuracy speaker verification systems are realizable for handling speech obtained under unspecified noise conditions.

ACKNOWLEDGEMENTS

My thanks and sincere appreciation is directed especially to my supervisor, Dr. M. Shridhar for his guidance, support and constant encouragement. The advice of Dr. G.A. Jullien and Dr. W.C. Miller is gratefully acknowledged. The help of the other members of the Electrical Engineering Department is fully appreciated.

I would like to express my sincerest thanks and gratitude to my wife, Anna, for her indulgence, help and constant encouragement. To my mother, I extend my sincere thanks, for her constant moral support.

Thanks are also due to Mrs. Sherry Sweeney, who did an excellent job of typing this dissertation.

TABLE OF CONTENTS

Table of Contents (cont'd)

Table of Contents (cont'd)

LIST OF TABLES

LIST OF ILLUSTRATIONS

TABLE OF APPENDICES

CHAPTER I

INTRODUCTION

The ability of the human ear to correctly establish the identity of an acquaintance by merely listening to the person's speech has long fascinated scientists and engineers. The importance of this speaker recognition capability of the human ear can be easily understood if one considers the following applications:

1) release of classified security information over telephone to a person whose identity is confirmed on the basis of listening tests
2) criminal investigations where a person's identity is established on the basis of speech rather than sight.

The above applications have prompted researchers to investigate the feasibility of developing an efficient scheme or a machine that can be programmed to conduct speaker recognition tests by analyzing a person's voice.

The successful development of such a device opens up a vast area of technological applications both in the industrial and consumer market. It is easy to visualize a situation where a customer conducts his banking business from his home or office, a resident having his garage door respond only to his voice or where restricted information is released to a person only after his identity is confirmed by an analysis of his spoken utterance.

At the present state of the computer technology it seems quite feasible to realize efficient, low cost recognition devices that have error rates less than one percent.

1.1 SPEAKER RECOGNITION

Speaker recognition can be broadly defined as the establishment or verification of the identity of a person based on an analysis of his spoken utterance. The problem of speaker recognition can be divided into two classes:

a) Speaker Identification

Speaker identification defines the task of selecting from among a finite population of N speakers that person with such voice characteristics that are most similar to those of the person whose identity is being claimed. It is easy to observe that if the unknown speaker does not belong to the finite population, a false identification could result. The overall probability of an incorrect decision is a monotonically increasing function of population size N.

b) Speaker Verification

Speaker verification deals with the acceptance or rejection of a person's claim of identity based on analysis of the unknown speaker's utterance. In speaker verification case, a single comparison of the unknown speaker's voice pattern with that of the person whose identity is being claimed is sufficient to either accept or reject the claim. Therefore, the probability of the false decision is generally independent

of the number of speakers. Much of the theory and techniques of speaker verification is common to speaker identification.

Once defined the speaker verification problem may be split into two groups:

i) text-dependent speaker verification system

ii) text-independent speaker verification system

The first group deals with a priori prescribed text, usually a number of utterances with a few seconds of duration or single words. The code phrases have been constructed to reflect certain combinations of vowels, voiced and unvoiced fricatives, plosive sounds which would be helpful for verification purpose. The second group is based on the fact, that the texts of the test and reference speech samples are different. Text independent speaker verification systems usually have some restriction on the nature of the speech material employed to the verification task; such as linguistic content and duration of the spoken text.

One of the assumptions in most speaker verification systems (SVS), both text-dependent and text-independent, is that speakers within the customer set are cooperative. It means that they do not intend to make any attempts to change their usual speaking behavior from one trial to another.

For many applications the potential customers are expected to be cooperative so that a prescribed text is perfectly feasible. The most investigations have been directed to text-

dependent systems and they are the closest to practical implementation in a real-world speaker verification mode.

Potential applications of SVS are in banking, voice authorized credit cards, entry control to restricted areas e.g., computer centers, criminology, etc.

1.2 SPEECH PROCESS

A great deal of speech investigations have studied the human speech mechanism to develop an appropriate speech production model.

In order to understand more about the speech process different models were postulated and tested for various conditions. An ideal desirable model would be a linear and time-invariant system. Unfortunately, the speech is a continually time varying process and does not exactly satisfy any of these conditions. However, with reasonable assumptions one can develop linear time-invariant model valid over short periods of time that describe the significant speech characteristics. The most frequently used speech parameters that have been investigated are fundamental frequency, voice intensity, resonant frequencies (formants) and spectral data.

In the last few years a concentrated effort has been directed towards replacing some of the above parameters by linear prediction coefficients and their transform equivalents.

Research has been motivated by the desire to produce an artificial speech which would sound as natural as possible.

Using different parameters to represent and encode the speech signal, a number of voice coders have been developed that generate the synthetic speech.

Vocoders offer not only quite intelligible speech but reduce the required bandwidth for transmission." Unfortunately, the naturalness of synthesized sound is rather poor and recognizability of a talking person is far from desired.

A natural problem has arisen how to distinguish among speech parameters to indicate features responsible for linguistic information and those related to the speaker himself.

Speech process is a result of a complex sequence of transformations occurring at several different levels: linguistic, semantic, articulatory and acoustic. Fluctuations of these transformations are related in general, to the fluctuations of the acoustic properties of the human speech.

Some variations in speech are due to the structural differences in the shape and length of the vocal tract and can be concerned as speaker dependent. The speaking behaviour, accent and voice melody are the examples of a group of learned features characterizing the talker.

Computerized speech analysis and synthesis have provided an acceleration to research in automatic speech recognition and speaker recognition areas. Ideally, the parametric representation of speech should be not only high informative but

reasonably easy to be simulated on general purpose digital computer or implemented in special hardware devices.

1.3 REVIEW OF SPEAKER RECOGNITION TECHNIQUES

Both speaker verification and identification systems have been investigated in past experimental studies. Visual analysis of spectrograms and analysis of utterances by a group of common listeners were among the oldest techniques [23] [24] for discriminating among speakers. Verification accuracy varied from 5 to 25% and depended on particular subjective skills of the verifying operators.

Efforts have been directed towards finding an automatic procedure in order to replace the human expertise in verification process by far more objective computer expertise.

One of the first methods [25] of automatic speaker recognition used the short-time power spectrum of speech as a speaker-specific feature.

In this experiment, the spectrum analysis was performed by the use of a 17-channel filter bank scanning the bandwidth of 100-7000Hz. Four repetitions of ten words spoken by ten speakers (7 male and 3 female) were used as speech material. A three dimensional array describing the spectrum as a function of frequency and time represented each analyzed word. The time alignment of the words was performed by lining up the peaks of the energy-versus-time function of each word. The reference pattern for each speaker was created by combining

three utterances of each of the ten words. The recognition procedure consisted of cross correlating the spectrographic pattern of the test utterance of each word with each of the ten reference patterns for that word. The highest correlation indicated the speaker of the test utterance. There were 393 cases tested and recognition accuracy was 89 percent. The recognition score was not uniformly distributed for all of the words and ranged from 74 to 97 percent. Accuracy of recognition for individual talkers ranged from 77 to 98 percent. Next step in this procedure was reducing the three-dimensional time-frequency-intensity patterns to two dimensions by averaging over time for each of the 17 frequency bands. Even with time dimension eliminated the recognition score of 89 percent was maintained. However, when the three-dimensional patterns were reduced to two dimension, the recognition accuracy decreased to 47 percent that implies that spectral information is more important than the energy time information for speaker recognition purpose.

Studies performed by Bricker, et al [26] on the same set of data resulted in 97 percent accuracy of recognition for two-dimensional spectral data.

They used the non-Euclidian distance measure which reduced the error rate by a factor of 4 as compared to the simple cross correlation measure. A 20-channel filter bank covering the frequency range 20-2900Hz was employed to obtain the spectral data.

Both studies demonstrated the importance of spectral information in speech for speaker recognition. However, there are some problems connected with the use of time-averaged spectra for practical systems. Since the spectral data as used in these studies is significantly influenced by the frequency characteristics of the recording and the transmission apparatus, any variations in these characteristics would be an additional source of randomness in the spectral data. This could cause an adverse influence on the performance of the overall system.

Another study based on spectral representation of speech obtained from a filter bank of 15 channels with frequency range of 300-4000 Hz was performed by Li et al [29].

Three test utterances were used in the experiment. "My name is .... (name)", "I.D. Reference (name)" and "Alibi". Recordings were made over telephone lines. Two-level adaptive linear threshold element system was employed to perform the speaker verification. At first level, a set of speaker dependent weights are determined from the bank channels and time segments. The decision making process takes place at the second level. The overall system accuracy is 90 percent.

The pitch or the fundamental frequency of vocal cord vibrations has been found to be an effective speech parameter for automatic speaker verification. [5] [12] [27] [28]. Pitch

has an important advantage over the spectral information: it is not affected by frequency characteristics of the recording and the transmission system. Atal [27] described the speaker recognition system based exclusively on pitch as a speaker-specific feature. The pitch extraction procedure was performed using a short-time correlation analysis of the cubed and low-pass filtered to 1kHz speech waveform. The pitch period was determined as the time interval corresponding to the largest peak in the short-time correlation function. The speech data consisted of six repetitions of the sentence "May we all learn a yellow lion roar". Ten female speakers participated in the experiment. The speech recordings were made in anechoic chamber using high-quality microphone on two different days 27 days apart.

Five utterances of each speaker were used to form his reference pattern while the sixth one served as the test pattern. The non-Euclidian distance measure was employed to discriminate speakers. The overall accuracy of identification was found to be 97 percent based on total of 60 judgments. The use of Euclidian distance measure gave only 68 percent accuracy.

An approach similar to Li's is the system described by Das and Mohn [30].

Reported system operates on signals from a filter bank but in addition to band energies, it includes pitch and formant

information. An error of about 1 percent is obtained but there is 20 percent "no decision" rate.

An automatic speaker verification system has been under investigation at Bell Laboratories. The first implementation was performed by Doddington in a large-scale Honeywell computer [5] [12]. The analysis involved the first three formants, intensity and pitch information. A population of 40 male speakers participated in the experiment. They all voiced utterance "We were away a year ago" was recorded as a test text. Speakers were divided into two groups: eight of them were "customers" and thirty two were "impostors". The equal-error criterion was employed in a system performance evaluation.

The average verification accuracy was around 98.5 percent. Doddington reported that the formant information contributed relatively little to the overall system performance. In addition, the formants analysis consumed approximately 200-300 times the duration of the test utterance.

An important modification of the Doddington system was performed by Lummis [4]. He tried to eliminate the time consuming formant analysis and replace the time registration previously controlled by second formant into the intensity controlled temporal registration. He used the same panel of speakers and the same test utterances as Doddington did. Average equal-error rate was reported at 1 percent when all the features were used in the distance computations and 1.2 percent when the

formants were eliminated. These results confirmed Doddington's observations about little formants contribution into speaker verification system performance.

However, in an additional experiment, Lummis and Rosenberg [31] reported that the information provided by formants is of value in the case of impostors like professional mimics who attempted to imitate customer utterances. Four professional mimics were employed to imitate the voices of eight customers. Following an intensive training the mimic utterances were recorded and processed. The dissimilarity measure based only on pitch and intensity resulted in 41 percent rate of acceptance of these utterances. Including three formants the false acceptance rate decreased significantly to 27 percent.

Pitch and intensity contents were also subjects of evaluation in an automatic speaker verification system implemented by Rosenberg [11]. Customers assessed the system via dialed-up telephone lines. The speakers population exceeded 100. They called up nominally once each working day over five months. The verification operation was performed on the base of the one, fixed sentence "We were away a year ago." The system was implemented in software on NOVA 800 Laboratory computer. Data set hookup enabled the telphone line access to the computer. The input signal was low-pass filtered with 900Hz cut-off frequency, digitized with 10kHz and 12-bit resolution, and stored on disk. Reference data was computed

off-line and updated with the analysis data of accepted utterances. The time registration based on dynamic programming technique was used. The sample intensity contour was linearly stretched or compressed to the normalized length of the reference intensity contour. The error rate of 10 percent was obtained for new customers and approximately 5 percent for stable and well established customers.

Effectiveness of linear prediction parameters in speaker verification operation was reported by Atal [8]. He investigated linear predictor coefficients and their transform coefficients as autocorrelation function, impulse response, the area function and the cepstrum function. The test population consisted of ten female speakers. They repeated six times the sentence "May we all learn-a yellow Lion roar". The cepstrum function was found the most effective speaker-dependent feature from among all the parameters investigated. The error rate of 17 percent was achieved by the analysis of the 50ms of speech and decreased to 2 percent for speech duration of one second. In the text-independent speaker identification experiment, Atal reported an error rate of 7 percent for speech duration of two seconds.

An entry control system using voice verification is in use at Texas Instruments as reported by Doddington [32]. Over 200 users are authorized entry and are enrolled on the speaker verification system. Women are 13 percent of the speakers

population. A daily average of 400 entries is made. The verification is based upon a four-word utterance. The words are randomly selected from a set of sixteen monosyllabic words. Four-phrase sequential decision strategy is used which makes it possible that the speaker can be accepted after the first phrase whenever the result of comparison is less than certain threshold. Otherwise, an additional phrase is requested combining the decision functions at each stage. Verification accuracy was found to increase by increasing number of utterances.

The customer reject rate was 13 percent at 1 percent impostor acceptance level when only one utterance was tested. A dramatic improvement up to 0.3 percent was reported when 4 utterances were used in verification procedures. It is worthy to mention, that speakers are verified in acoustic booth and use dynamic microphone. Speech reference data and speakers parameters are adaptively updated each time the speaker successfully verifies his identity.

In recent experiments [10] Sambur reported a verification system without the need of any time normalization procedure. The speaker discrimination potential of the linear prediction orthogonal parameters was formally tested in both a speaker identification and verification scheme. The speaker recognition experiment involved 21 male speakers. The speech data for the text-independent study consisted of six repetitions of the same

utterance spoken by each speaker. The sentence analyzed was "I was stunned by the beauty of the view" The recording sessions were made in a quiet environment and were spaced over three weeks. The speech was bandpass filtered from 100Hz to 4kHz and digitized with 10kHz sampling rate. The utterances were divided into frames with fixed duration of 20 ms. A 12th order of linear prediction analysis was employed to obtain linear prediction parameters, parcor coefficients and area coefficients. The recognition accuracy of 99 percent for high quality speech signals was reported. For telephone input the accuracy was 96 percent. The total results of text-independent speaker recognition study showed the accuracy near 94 percent.

Another evaluation of speaker verification system in Bell Laboratories was undertaken by Rosenberg and Sambur. [21]. This improved system has been augmented to include linear prediction parameters in addition to the already existing pitch and intensity analysis of sentence-long utterances. A method for selecting optimum speaker-dependent features has been incorporated in this system. The goal of this study was to investigate new features for analysis to supplement pitch and intensity parameters, replace time consuming formant analysis, and improvement of system performance particularly with respect to mimic utterances.

Four speakers participated in this experiment. All voiced

utterances "We were away a year ago" and "I know when my lawyer is due" and these were chosen as a test sentence. The 12th order of linear prediction was used to perform analysis on fixed duration frames 10ms. Finally a set of two prediction coefficients a_4 and a_8, pitch and gain were employed as speaker-specific features and corresponding contours were created. The evaluation showed, that overall verification error was approximately 1 percent for casual impostors and 4 percent for well-trained mimics.

It indicated the significant improvement of the verification accuracy over previous implementations which used only pitch and intensity, even with formant analysis included.

Effectiveness of low frequency parameters for speaker verification was reported by Vidalon [40]. The speech was low-pass filtered to 2kHz and the 6th order of linear prediction model was applied. Six speakers participated in the experiment. Using composite references and a set of selected parameters, namely the 6th linear prediction coefficient and the 7th autocorrelation coefficient, the error rate 2-3% was reported. The speech utterances were recorded in anechoic chamber.

A conclusion of this survey is, that there still exists gaps in our knowledge with respect to speaker recognition. Especially, requirements for a reliable-with accuracy better than 95 percent - speaker verification system evaluated over dialed-up telephone lines are of great interest.

Our investigations will be directed toward solving this particular problem.

1.4 THESIS MOTIVATION

Most of the proposed speaker verification systems have dealt with the speech utterances recorded under relatively noise free conditions. Some of these techniques have reached the sufficient degree of verification accuracy to be implemented in a reliable SVS. However, there are several aspects of the speaker verification which deserve further attention. There are practical limitations of systems based on high quality speech signals.

At the present time, applications of the greatest interest are the use of voice signal for remote personal verification performed on telecommunications channels operating under "real world" conditions.

Different types of noise, distortions and reduced bandwidth inherent with the signal, make verification over transmission channels a challenging problem to be solved.

This thesis considers some of the problems associated with practical design of a speaker verification system operating with different speech inputs: high quality, telephone and speech with additive wideband noise. The linear prediction model of speech is employed in this research. The linear prediction coefficients contain the combined information about

the spectrum of the speech signal, distribution of formants, their bandwidth and the glottal waveform.

Moreover, they are relatively easy to compute in time domain, directly from sampled signal using autocorrelation or covariance method.* In opposition to formants analysis which is quite complicated and time consuming operation (approximately 300-400 times the test utterance), the computation of LPC parameter is relatively fast.

Recent results [3] [10] have shown the potential of the orthogonal linear prediction parameters to differentiate among talkers and their capability for speech synthesis purposes.

In this dissertation the speaker-specific features of part of orthogonal parameters are confirmed. However, the simple time normalization procedure a priori to the LPC parameters extraction was found of significance to obtain reliable results of verification.

One of the objectives of this work is to investigate the sensitivity of orthogonal parameters to noise and the influence of the noise on their recognition potential.

A study is required in order to determine the relation between the linear prediction coefficients and the orthogonal parameters of a clean speech signal and their equivalents derived from noisy speech.

An adequate order of linear prediction model used to parametric representation of a speech with different signal-to-

*See Appendix B.

noise ratios is under consideration. Optimal order and number of speaker-dependent features tend to improve the overall SVS efficiency. Finally, the speaker verification system operating over telephone lines is presented. The thesis demonstrates the feasibility of the optional operation of SVS with high quality or telephone speech references.

1.5 ARRANGEMENT OF DISSERTATION

Chapter I contains the revue of available literature in the field of speaker recognition. Different techniques and results of experiments are submitted. Some principles of speech process are explained and problem definition is stated.

Chapter II describes speaker-dependent acoustic parameters of the human speech, their definitions and methods of computation. The significance of linear prediction parameters as a powerful tool in speaker recognition process is emphasized. Finally, the data base and processing methodology for speaker verification system implemented around minicomputer NOVA-840 are presented.

Chapter III describes the step-by-step algorithm for speaker verification. Important properties of some of mutually uncorrelated parameters derived from eigenvector analysis are discussed. Possibility of reduction of reference vector dimensionality is demonstrated. The necessity for time normalization of speech utterances is substantiated.

Problem of adequate distance measure and threshold of verification is discussed. Results of verification for high quality speech are presented.

Chapter IV contains the original study on sensitivity of orthogonal parameters as speaker-dependent features, to noise. Some mathematical relations are derived between linear prediction coefficients and their orthogonal parameters from noise added speech, and their corresponding parameters retrieved from clean speech. Results of verification procedures performed on speech with additive noise are presented. Optimal number of verification parameters in function of different signal-to-noise ratio and order of prediction is discussed.

Chapter V describes the speaker verification system over telephone lines. It presents some aspects of noise and distortions influence on verification ability of the system. Spectral analysis of a noise and signal in telephone channel are performed. Verification accuracy as a function of orthogonal parameters are reported and results are briefly discussed. Finally, the verification time is estimated.

Chapter VI presents the new concept of speaker verification system operating on noisy test utterances versus references obtained from high quality speech. Problems of adaptive noise cancelling are considered. Technique for removing the

noise from telephone channel is presented. Significance of speech enhancement to accuracy of verification is reported.

Chapter VII contains the main conclusions of the thesis devoted to speaker verification domain.

CHAPTER II

SPEAKER-SPECIFIC FEATURES EXTRACTION

2.1 INTRODUCTION

A crucial point in the success of any pattern recognition system is the proper selection of features that adequately characterize the patterns of interest.

In the last few years a great deal of studies have been undertaken to determine a set of acoustic features in the speech signal that are the most effective for speaker recognition purposes.

By "most effective" it is implied that the chosen parameters should represent the unique properties of a speaker's vocal tract and glottis as well as some aspects of his learned pattern of speaking. They should characterize the speaking rate, stress melody and coarticulation, to mention only the most important.

In the ideal case, they ought to contain no information at all about linguistic content of the speech.

In practice, unfortunately, it is almost impossible to extract purely speaker identification features, that wouldn't be deteriorated by additional linguistic information.

In the following section, different speaker-dependent features and techniques of their extraction will be described.

2.2 ACOUSTIC PARAMETERS OF SPEECH

Speech sounds are produced as a result of acoustical excitation of the vocal tract which consists of the cavities in pharynx and the mouth. Overall speech production model consists of three main components: glottis, vocal tract and lips. The more detailed description of speech model is contained in Appendix A.

In a simplified model of speech production the vocal tract can be represented as a linear time-varying filter excited either by periodic pulses or by a pseudo-random noise source. Voiced sounds are generated when periodic pulses are applied to the input of the vocal tract. Unvoiced sounds are produced when the vocal tract is excited by a noise-like turbulent flow of air at a point of constriction.

Both kinds of sounds contain much information about spoken message and the speaker himself. The extraction of information which reflects fixed anatomical properties of the vocal tract is the fundamental problem in automatic speaker verification system.

The acoustic parameters of speech which have been found the most effective for speaker recognition are presented below:

a) gain or intensity -- is a function of time defined over a period T of range 10 - 30 ms.

$$I(t) = \int_{t-T/2}^{t+T/2} x^2(t)\,dt$$

The change of voice intensity is a result of variations of the vocal tract shape and the subglottal pressure. The gain parameter represents an important speaker-specific feature. [4] [12].

b) pitch -- is a fundamental resonance frequency of the vocal cords vibrations.

There are several methods of pitch period estimation. Peak picking of autocorrelation of signal or error signal and average magnitude difference function (AMDF) [37] [38] are the most frequently used methods in time domain. Computing the frequency spacing of the spectral peaks is the method used in frequency domain (SIFT) [1].

c) formants -- are the resonance frequencies of the vocal tract. Usually, first three formants which occur in the range 0-3kHz of human voice are of main interest in speaker recognition tasks. They are very speaker-specific features. Although there exist several methods of formants estimation and their bandwidths [1] [33] [34], there are still problems with accurate and reliable determination of formants for both male and female speakers.

Moreover, the computation time exceeds by 200-300 times the speech duration. This is excluding the format analysis from the on-line speech processing.

d) short-time spectrum -- gives a three-dimensional representation of the speech signal in terms of energy, time and frequency.

The short-time power spectrum is defined as

$$G(f,t) = \left| \int_{-\infty}^{\infty} s(T)w(t-T) \exp(-j2\pi fT)dT \right|^2$$

where $s(t)$ - voice signal

$w(t)$ - appropriate window function

The length of window function is usually 20-30ms. The Hamming window is frequently used in short-time spectrum computation.

There are two methods of calculation.

1) In one method, the short-time spectrum is obtained from a bank of band-pass filters.

Wolf [28] applied a 36-channel bank, scanning linearly a band between 150Hz and 1650Hz, and logarithmically over 1650Hz up to 7025Hz.

2) The second method employs the Fast Fourier Transform and digital computer to power spectrum estimation [1].

Since the short-time speech spectrum contains almost complete information about acoustical characteristics of human voice, it has been found to be effective in speaker recognition domain.

e) spectral correlations -- provides another factor enabling to discriminate among speakers.

There is a significant correlation between the short-time spectrum at different frequencies. Spectral correlations vary consistently from one person to another [35]. However, to evaluate reliable correlations, the averaging process must be taken over long period of speech, at least of 30 sec. This approach may be useful for text-independent speaker verification systems.

f) linear prediction coefficients (LPC) -- are the results of linear prediction analysis which involves the theory of prediction in speech analysis. The theory says, that within a time of 10-30ms, when the speech process is assumed to be stationary, the actual speech sample can be predicted as a linear weighted sum of the previous samples. These weights, LPC's in time domain, and recursive filter coefficients in frequency domain -- are estimated by minimizing the mean-squared error of prediction between the true value of the signal and predicted one.

The LPC parameters represent the tremendous amount of combined information about short-time power spectrum, formants and their bandwidths.

Moreover, they are relatively easy and fast to compute on digital computers using several techniques such as autocorrelation or covariance methods, described in Appendix B. The LPC parameters can be directly transformed into log area coefficients, or reflection (Parcor) coefficients.

All of them have been found to be very effective in speaker recognition investigations [8] [21] [36].

g) nasal coarticulation -- is a result of low movement of the articulators in connected speech.

The shape of the vocal tract at given time depends not only on the phoneme being spoken actually, but also on the neighboring phonemes. The coarticulation in a given context is speaker dependent. Studies by Su et al [22] have shown that coarticulation during the production of nasal consonants is useful for speaker recognition. In this experiment, an acoustic measurement of nasal coarticulation was performed by computing the spectral difference between the mean spectrum of a nasal consonant followed by a front vowel as "i" and that of the same consonant followed by a back vowel as "a".

2.3 SPEECH DATA PREPARATION

Utterances to be used in a speaker verification experiment should include a wide variety of speech sounds; vowels, fricatives, stops, nasals and diphthongs. They should also

be devised in such a way that they are easy to segment, natural to say, and usually spoken in just one way. Based on these considerations a set of following sentences were used in our experimentation.

1) We were away a year ago.
2) I know when my lawyer is due.
3) Papa needs two singers.
4) Pay the man first, please.
5) My name is Miller; cash this bond, please.
6) I was stunned by the beauty of the view.

First two phrases are all voiced sentences frequently used in speaker recognition studies. It is proven, that an individual's voice may change from day to day. In some of the early works on speaker recognition, both the test and reference utterances were recorded in a single session.

Later on, it became obvious that the results from such studies can be misleading. A speaker can often maintain a high degree of stability in his speaking habits over a short interval of time. However, the intraspeaker variability can often increase significantly with an increase in the time interval between recordings of utterances. Therefore, in our experiment, the data were collected over a period of time.

There were six recording sessions spaced over 7-8 days during the period of 45 days. Recordings were made in quiet

room conditions. Six sessions of high quality speech over a dynamic microphone (with 40-16,000Hz frequency response) and six sessions over dialed-up telephone lines were recorded.

Telephone calls were initialized from different places within University of Windsor or from the speaker's homes. There were four sessions recorded over local telephone exchange and two sessions over the city network.

During second and third sessions the test customers were asked to speak simultaneously to the telephone and dynamic microphones and the operator recorded both versions of a speech on the dual-tape Hi-Fi tape recorder in the remote laboratory room.

The reason to make these parallel recordings was to enable the reliable, precise comparison of high quality speech with noisy and distorted telephone signal (SNR, spectrum).

During first five sessions, the participants were asked to repeat six times a whole set of test utterances. However, in the sixth session, they repeated in turn, each utterance six times.

It should be underlined, that speakers were not trained, in any manner and did not imitate anyone. They were asked to utter the test phrases as naturally as possible. However, after careful listening to the recorded sentences in all

sessions, a kind of routine-like speaking behaviour was observed during the sixth session. It has confirmed our previous suspicions, that multirepetition of the same utterance introduces the undesirable effects of "List intonation" which alters the speaking habit into the less natural.

Ideally, the data base for speaker-verification studies should include a large number of speakers. Collection of speech data from a large number of talkers poses many practical problems especially in small mini-computer oriented laboratory. Consequently, most studies use a relatively small population of speakers. With a small number of speakers, it is preferable not to include persons with widely different speech characteristics.

For example, male and female voices can often be distinguished from each other on the basis of average pitch alone.

For the purpose of our experiment, a group of six male speakers was chosen. All of them were within age 30-40, were native Canadians from Windsor area and had no noticeable differences in accent.

The recorded speech was band-pass filtered to 200-3200Hz, which conforms to the telephone bandwidth and then digitized using analogue-to-digital converter with sampling rate of 8kHz and 14-bit resolution. The digitized speech was stored in disk for further processing.

2.4 EXPERIMENTAL SET-UP

A functional block diagram of the analog and digital processing of our speaker verification system is shown in Fig. 2.1.

The system is built around Central Processor Unit of a NOVA-840 minicomputer. The analog processing group contains; Hi-Fi 4-channel taperecorder - Crown -SX800 with remote control, Krohn-Hite variable filter, Model 3750, for range 0.02Hz to 20kHz, with four optional attenuation slopes of 6, 12, 18 or 24dB per octave, differential amplifier NEFF-Type 119 used for recordings from the telephone line, and telephone hand-set.

The digital processing part contains:- a NOVA 840 minicomputer with the memory of 32K words, accompanied with two Diablo disks with combined capacity of 2.5 million words, Tustin Electronics Analog-to-Digital converter with 14-bit resolution, possible sampling rate up to 100kHz and maximum dynamic of input signals $\pm$ 10V, Tektronix-4013 Graphics Terminal to display processed waveforms, Tektronix-4610 Hard Copy unit, and Infoton Vistar CRT terminal used to control the operations of computer.

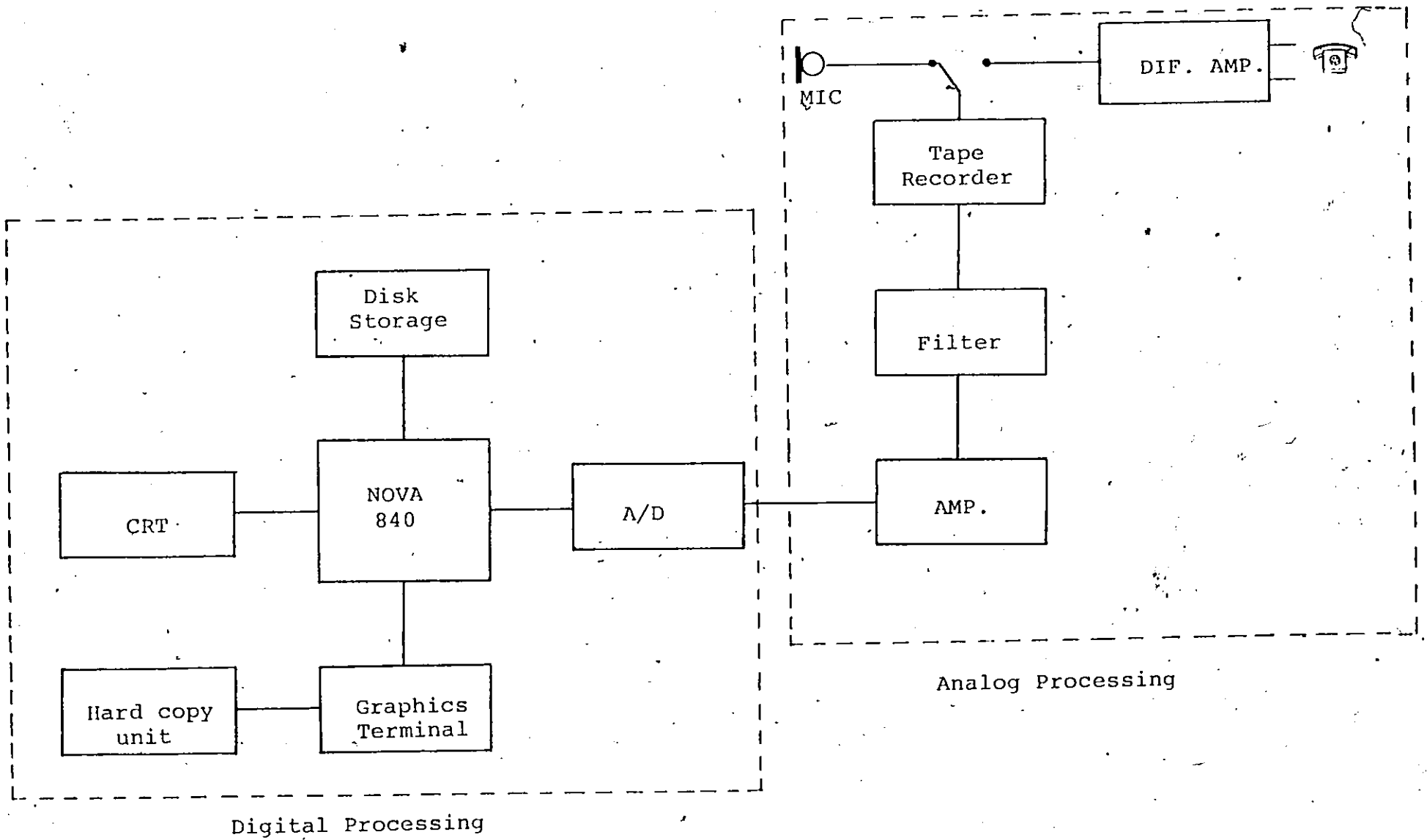


Fig. 2.1. Block Diagram of Experimental Set-Up

CHAPTER III

SPEAKER VERIFICATION ALGORITHM

3.1 INTRODUCTION

The proposed speaker verification system deals with an a priori arranged set of test utterances. The person to be verified is asked to utter the prescribed phrase. Sometimes the additional repetition or second phrase is required. It is assumed that speakers are cooperative and speak with their own, natural behaviour; in other words, they wish to be verified positively. The utterances have been chosen to be advantageous for the Speaker Verification System and there is no way to omit their repetition to be verified.

It is so called "text-dependent" SVS.

The verification algorithm essentially requires only one comparison. Occasionally, in a case when the result of verification lies within an "uncertainty" or "lack of decision" margin -- the talker, who claims his identity is asked to repeat the same sentence again or to speak another sentence from the prearranged set of phrases.

The speech signal processing procedure is then applied to measure and estimate pattern parameters that are as much as possible indicative about speaker's identity. The pattern parameters are compared to the reference pattern, already created and stored in the computer memory.

Pattern comparison and decision making process is the final step of the algorithm which generates the answer: accept or reject the identity claim.

There are many applications for an automatic speaker verification system. To mention only: the voice banking, the voice validation of credit cards, keeping the guard to the restricted areas e.g., data bank in computer centres, criminology, etc.

Figure 3.1 illustrates the basic block diagram of presented SVS.

3.2 LPC ANALYSIS

In this experiment, the linear prediction coefficients were used as primary speaker sensitive parameters. The LPC parameters can be estimated using several methods like: autocorrelation, covariance, lattice structure and others.

The autocorrelation method which assumes the stability of the recursive prediction filter [1] was used in this LPC analysis.

The description of the method the reader can find in Appendix B.

The LPC parameters are computed on a frame by frame basis. The frame duration is not fixed and varies from utterance to utterance, but it is not longer than 25ms to assume the statistical properties of the vocal tract are constant within frame time. The speech data is bandlimited

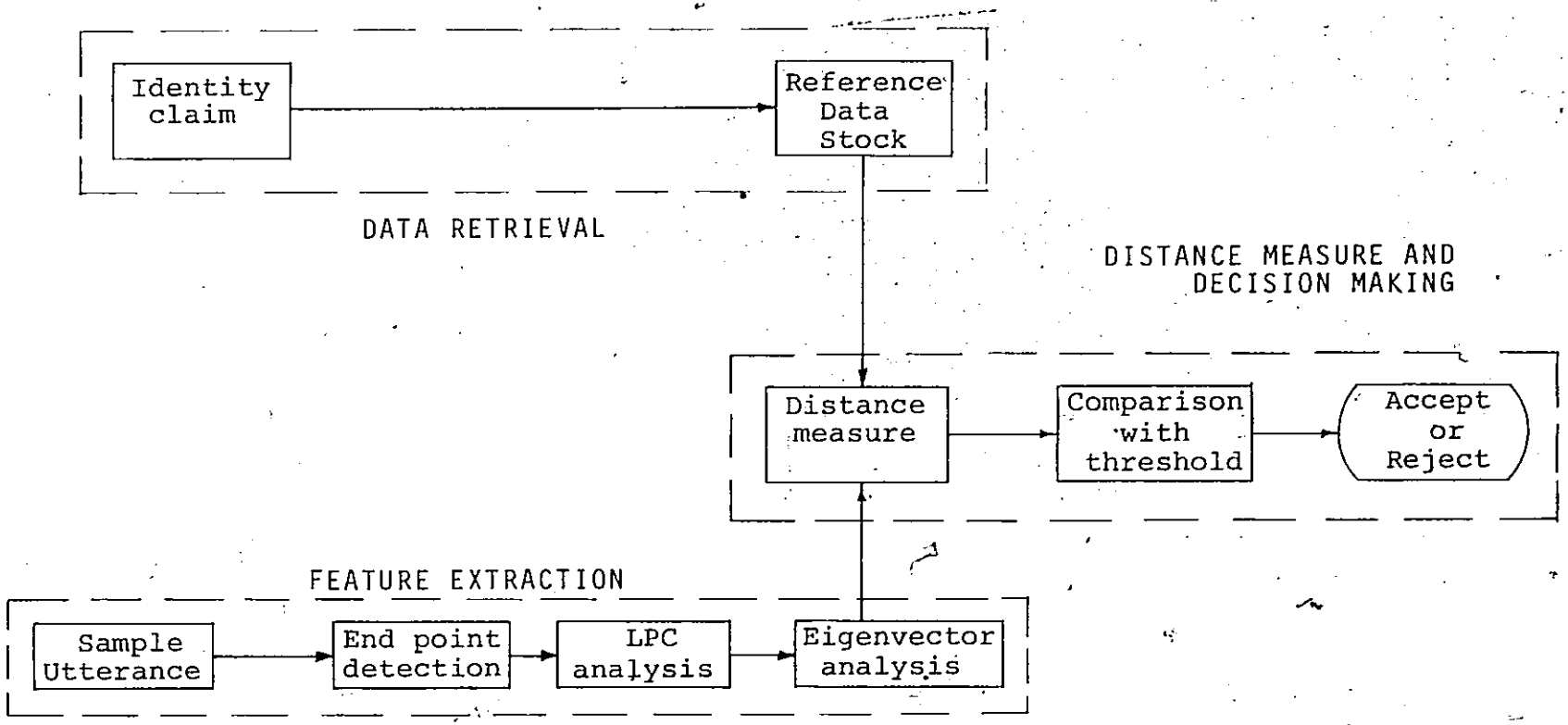


Fig. 3.1. Speaker Verification System -- operation block diagram.

to the 200Hz ÷ 3200 Hz band, and sampled at a frequency of 8kHz.

At least the first three formants lie within this band, which seems to be sufficient for proper analysis of the speech waveform.

Three orders of linear prediction; 8th, 12th, and 16th were applied during our experiment to find out their influence on the speaker verification

3.3 EIGENVECTOR ANALYSIS

It is known, that the linear prediction coefficients used to model the speech production are redundant i.e., they carry much more information than is needed to synthesize the speech [2].

To reduce the redundancy of LPC, a new vector of parameters can be created by using the eigenvector analysis [3].

Transformation of LPC parameters into a set of orthogonal eigenvectors requires the creation of the covariance matrix C across the total utterance. The elements of the matrix are defined as:

$$(3.3.1) \qquad C_{ik} = \frac{1}{NF-1} \sum_{j=1}^{NF} \left(a_{ij} - \bar{a}_i\right)\left(a_{kj} - \bar{a}_k\right)$$

where a_{ij} - the i-th LPC in the jth frame

NF - number of frames

$\bar{a}_i = \frac{1}{NF} \sum_{j=1}^{NF} a_{ij}$ - average value of the ith LPC

Since the matrix C is real and symmetric.

(3.3.2) $$C_{ik} = C_{ki}$$

$$\text{where } \begin{matrix} i \\ K \end{matrix} = \left.\begin{matrix} \\ \\ \end{matrix}\right\} 1, 2 \ldots\ldots N$$

N -- order of prediction

hence only N(N+1)/2 elements need to be computed. Real symmetric matrices have two fundamental properties.

a) eigenvalues are real

b) eigenvectors form an orthogonal set

If the matrix C has N distinct eigenvalues, λ_i, the N resulting eigenvectors v_i are linearly independent i.e., they are mutually uncorrelated.

One can express it in the following form:

(3.3.3) If $\lambda_1 \neq \lambda_2 \neq \lambda_3 \neq \ldots\ldots \neq \lambda_N$

then $\langle v_1, v_2, v_3 \ldots\ldots v_n \rangle = 0$

The eigenvectors v_i of the matrix C can be determined solving a set of N-simultaneous linear equations.

$$
\begin{aligned}
&(C_{11}-\lambda)v_1+C_{12}v_2 + \ldots\ldots + C_{1N}v_N = 0 \\
&C_{21}v_1 + (C_{22}-\lambda)v_2 + \ldots. + C_{2N}v_N = 0 \\
&\qquad\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&C_{N1}v_1 + C_{N2}v_2 + \ldots\ldots\ldots + (C_{NN}-\lambda)v_N = 0
\end{aligned}
\tag{3.3.4}
$$

In the matrix form

$$[C - \Lambda I]V = 0 \tag{3.3.5}$$

where v_i - vector with N elements

I - identity matrix

The system of above equations has a nontrivial solution if, and only if, the determinant vanishes, i.e., if

$$\left|C-\Lambda I\right| = 0 \tag{3.3.6}$$

which results in a set of N eigenvalues

$$(\lambda_1, \lambda_2, \ldots.\lambda_N)$$

3.4 ORTHOGONALIZATION OF SPEAKER'S IDENTITY PARAMETERS

The linearly independent i.e., uncorrelated parameters expressed in terms of linear prediction coefficients will be called the orthogonal parameters.

The i^{th} orthogonal parameter in the j^{th} frame is defined as:

$$(3.4.1) \qquad P_{ij} \triangleq \sum_{k=1}^{N} v_{ki} a_{kj}$$

$$\text{where } \begin{matrix} i \\ k \end{matrix} = \Big\} 1, 2, \ldots N \qquad j = 1, 2, \ldots NF$$

In the matrix formulation the vector of the orthogonal parameters in the particular frame is given by

$$(3.4.2) \qquad \psi = A^T V$$

where A^T - vector transpose

The mean value of the i^{th} orthogonal parameters over the whole test utterance is computed from the formula

$$(3.4.3) \qquad \bar{P}_i = \frac{1}{NF} \sum_{j=1}^{NF} P_{ij}$$

Eigenvalues of the orthogonal parameters calculated across the utterance,

"WE WERE AWAY A YEAR AGO."

are demonstrated in the Table 3.1. Two orders of linear prediction, 8^{th} and 12^{th} were used in speech modeling.

As one can observe the high positioned eigenvalues are relatively small in both cases. It means, that corresponding orthogonal parameters are practically constant during spoken test speech.

	N = 8
1	24.71
2	6.648
3	1.0327
4	0.4141
5	0.1198
6	1.363×10^{-2}
7	1.1941×10^{-3}
8	1.0888×10^{-4}

	N = 12
1	94.46
2	27.83
3	9.078
4	3.218
5	1.1103
6	0.3861
7	0.1875
8	5.965×10^{-2}
9	1.366×10^{-2}
10	2.234×10^{-3}
11	8.401×10^{-4}
12	1.655×10^{-4}

Table 3.1 Eigenvalues of the orthogonal parameters calculated across the utterance

"WE WERE AWAY A YEAR AGO"

for two orders; 8 and 12 of linear prediction

The low positioned orthogonal parameters which indicate the significant standard deviation are assumed carrying the linguistic content of a speech and can be utilized to speech synthesis, efficient encoding and low bit rate digital signal transmission. (<1500 bit/sec) [3].

However, the least significant almost constant parameters, seem to carry some information about the particular speaking behaviour of a person, which is assumed constant during the spoken phrase.

The results of an experiment with two speakers speaking the same phrase are demonstrated in Tables 3.2 and 3.3, where the corresponding orthogonal parameters are shown. Again two linear prediction models were used.

The speaker A was asked to speak twice the same phrase.

The results obtained indicated, that for an 8^{th} order model the last four, and for a 12^{th} order model the last five orthogonal parameters were quite similar for both utterances of speaker A.

However, the corresponding orthogonal parameters for speaker B significantly differ.

It has confirmed, that the least significant orthogonal parameters are potentially informative about speaker's identity. These observed parameters will be used to establish the speaker verification algorithm.

N=8	SPEAKER A utter I	SPEAKER A utter II	SPEAKER B
1	1.0991	-0.8251	0.6085
2	-1.2973	0,3971	-2.4334
3	-1.6053	1.3566	3.5483
4	2.4944	1.1074	-0.9871
5	-3.2445	-3.2442	-1.0035
6	1.2394	1.4923	0.1531
7	-0.9118	-0.9987	0.5594
8	1.1398	1.5167	-0.4125

Table 3.2 Orthogonal parameters of 2 speakers computed for the 8th order model

N=12	SPEAKER A utter I	utter II	SPEAKER B
1	-0.1665	0.8657	-0.6921
2	0.2247	-0.5643	2.8707
3	-0.2801	-2.1393	-0.2838
4	-0.6371	0.2607	-1.1386
5	0.9114	3.7511	03.6419
6	-1.6578	-0.1229	02.3993
7	1.7523	-2.2979	02.1684
8	-1.6145	-2.0252	02.6974
9	1.8566	1.9271	02.0757
10	-0.9471	-0.7707	-2.0007
11	0.5670	0.4131	1.5932
12	-0.5597	-0.6337	2.8786

Table 3.3 Orthogonal parameters of 2 speakers computed for the 12th order model

3.5 CREATION OF REFERENCE PARAMETERS

In this experiment the six repetitions of an utterance for a given speaker was divided into a reference set consisting of five utterances and a test set. The reference set is destined to evaluate the individual reference pattern. Each utterance in turn is within the test set. Thus, for a given speaker six reference sets and six test sets were created.

The test set usually consists of one utterance of a reference speaker and the utterances of other speakers.

Each of six utterances for a given speaker were recorded on the separate session. To create the reference verification parameters first, one has to create the reference intraspeaker covariance matrix.

It can be performed by computing the mean value of L covariance matrices within the reference set and take it as a reference matrix C_{ref} expressed as

$$C_{ref} = \frac{1}{L} \sum_{l=1}^{L} C_l \qquad (3.5.1)$$

The reference eigenvalues are obtained from eq. (3.3.6) and reference set of eigenvectors V_{ref} from

$$\Lambda_{ref} V_{ref} = C_{ref} V_{ref} \qquad (3.5.2)$$

Then a group of orthogonal parameters calculated via reference eigenvectors is obtained for each of L utterances

of the reference set.

$$P_{ij} = \sum_{k=1}^{N} V_{ki_{ref}} a_{kj} \qquad (3.5.3)$$

The mean value of i^{th} orthogonal parameter $\bar{P}_i$ in every set of the reference utterances can be derived following the eq. (3.4.3).

Finally the set of reference orthogonal parameters are computed using the below expression

$$P_{i_{ref}} = \frac{1}{L} \sum_{l=1}^{L} \bar{P}_{il} \qquad (3.5.4)$$

The parameters thus obtained are then stored in the computer memory as reference pattern for the given talker and are then used for comparison and verification.

3.6 TIME NORMALIZATION

The main goal of this operation is to align corresponding speech events in the reference and test utterances. The proper time alignment allows the comparison of equivalent speech events and compensates the expected variations in the occurrence of these in repetitions of a given utterance.

There are several time registration techniques successfully implemented in the speech and speaker recognition [4], [5] and [6].

However, in our investigations, we deal with the speaker recognition features, which are the mean values of orthogonal parameters across the utterance or set of utterances. Hence, theoretically, the division of every utterance into number of frames with the fixed time duration T, where $10 \leq T \leq 30$ms [1] should be sufficient for speech segmentation. The number of frames would vary from say, 75 to 150 and no time normalization procedure would be required.

It was found nevertheless in our experiments, that the orthogonal parameters are not completely independent of the linguistic content of speech and that same kind of time normalization is essential in the speech data preprocessing. [See Chapter 3.9]

In the present implementation after carefully removing the silent portions from the beginning and the end of utterances, the simple time normalization was applied by dividing every utterance into a fixed number of frames. This number was chosen to be 100. The duration of frames varies from 12.5ms to 26ms which is within the range adequate to assume the stationarity of the speech process.

3.7 DISTANCE MEASURE

The degree of dissimilarity between two patterns is called "distance."

The measurements of distance between the reference pattern and the pattern for a given speaker (or speech segment) to be verified, has been investigated by researchers for years.

The single percentage error used sometimes as a distance measure is defined by

$$d = \frac{\sum_i (r_i - t_i)^2}{\sum_i r_i^2} \times 100\% \qquad (3.7.1)$$

where t_i - test vector

r_i - reference vector

i = 1, 2, 3, ... M-length of a vector [e.g. order of LPC]

The unweighted Euclidian distance measure is another frequently used method to compare the feature vectors. [7]

$$d_i^2 \triangleq (t_i - r_i)^T (t_i - r_i) \qquad (3.7.2)$$

A concept of a weighted Euclidian distance measure for speaker verification purpose first was investigated by Atal [8].

He applied this form of measure for linear prediction coefficients, the autocorrelation coefficients, the reflection coefficients, the impulse response of the filter 1/A(z),

and the cepstral coefficients of the inverse filter.

The distance is expressed as

$$d_i^2 \triangleq (t_i - r_i)^T C_{ref}^{-1} (t_i - r_i) \tag{3.7.3}$$

where C_{ref}^{-1} - the inverse intraspeaker reference covariance matrix (defined in the previous chapter).

Itakura [6] suggested the original method of distance measure used in frame by frame comparisons of linear prediction coefficients. This technique has applied the log likelihood ratio to compare two sets of LPCs.

$$d = \log \left(\frac{aRa^T}{bRb^T} \right) \tag{3.7.4}$$

where a's and b's are LPC parameters

R - reference correlation matrix defined as

$$R = \left[r(|i-j|) \right] \tag{3.7.5}$$

$$i,j = 0, 1, 2 \cdots, N$$

with elements

$$r(i) = \frac{1}{NS} \sum_{n=1}^{NS-|i|} s(n)s(n+i) \tag{3.7.6}$$

where s(n) - reference input signal

NS - number of analyzed samples in the frame

Another distance measure has been suggested by Cramer [9] and applied by Sambur [10].

If the pattern parameters are mutually statistically uncorrelated, the measure of dissimilarity can be expressed as a weighted distance with eigenvalues as weights i.e., $\lambda_1, \lambda_2 \cdots, \lambda_N$

$$(3.7.7) \qquad d = \sum_{i=1}^{N} \frac{(t_i - r_i)^2}{\lambda_i}$$

The above distance has been found to be very effective in dealing with orthogonal parameters which are mutually uncorrelated.

In our experiment, one can express eq. (3.7.7) as

$$(3.7.8) \qquad d = \sum_{i=1}^{N} \frac{(\bar{P}_i - P_{i_{ref}})^2}{\lambda_{i_{ref}}}$$

where $\bar{P}_i$ - orthogonal parameters of an imposter

$P_{i_{ref}}$ - reference orthogonal parameters

$\lambda_{i_{ref}}$ - reference eigenvalues

The proper choice of lower boundary of the summation (3.7.8) may not be 1, but that will be discussed in later sections.

3.8 THRESHOLD ESTABLISHMENT

One of the most important steps in speaker verification procedure is the determination of an appropriate threshold of recognition.

The specific point on the distance scale below that the test speaker is accepted as a true speaker and above that is rejected has been commonly named as a threshold of verification.

Using the distance measure given by equation (3.7.8), the distances were computed for the reference and test sets. In order to establish the proper threshold for acceptance or rejection of the identity of a given speaker, the following definition is proposed. [14].

The Minimum Rejection-to-Acceptance Ratio (MRAR) is defined as the ratio of the smallest distance from among the impostors, to the largest distance for a true speaker. MRAR may then be interpreted as representing the worst case in speaker verification for one speaker.

The largest distance D_{max} for a true speaker is determined by computing the distances associated with 5 utterances within reference set.

The maximum distance is then substituted as a divisor of MRAR.

In the experiment, 3 orders of linear prediction model have been tested to obtain an appropriate verification

threshold. Typical values of distance measurements computed for one reference spekaer WM are presented in the Table 3.4, 3.5, 3.6 . The reference orthogonal parameters have been derived from 1, 2, 3, 4 & 5^{th} utterances of the WM speaker.

Test phrase—"We were away a year ago."

It should be mentioned, that only the worst cases for every imposter i.e. representing the closest distances to the reference speaker have been shown in the table.

The distance obtained from the test utterance of the true speaker WM versus his own references has been also submitted into the table to have the idea, how distances differ between the true and false person. It can be observed that the largest MRAR was obtained when the least significant orthogonal parameters were used in the distance computation.

The extensive studies including all the speakers and 3 orders of linear prediction revealed, that the MRAR is greater than 2 in all cases. Therefore, a critical Acceptance Ratio is defined to be equal to 2.

In a practical test, the utterance of the test speaker is analyzed and the distance measure given by equation (3.8.8) is evaluated.

Thus the ratio of this distance to the largest distance D_{max} for the true speaker (previously computed and stored

SPEAKERS

$N^1 \rightarrow N$	WM6	JN3	FE5	JS1	JD3	PA3	D_{max}	MRAR
1 → 8	23.3	224.6	50.5	299.3	84.3	419.5	21.6	2.33
5 → 8	12.5	209.7	23.9	82.2	38.8	226.8	11.9	2.01
7 → 8	3.7	45.0	11.1	52.1	33.4	50.4	3.5	3.17

Table 3.4. Typical values of distances for reference speaker WM. The 8th order model. (N^1-1) - parameters excluded from computation.

SPEAKERS

$N^1 \rightarrow N$	WM6	JN6	FE5	JS6	JD1	PA1	D_{max}	MRAR
1→12	31.1	565.	139.6	507.	231.5	392.6	28.5	4.9
9→12	14.0	262.2	95.0	335.7	95.5	42.5	12.9	3.3
10→12	4.1	135.9	84.9	210.3	94.5	40.3	3.8	10.5

Table 3.5 Typical values of distances for reference speaker WM. The 12th order model.
(N^1-1) - parameters excluded from computation

SPEAKERS

$N^1 \rightarrow N$	WM6	JN6	FE5	JS6	JD1	PA1	D_{max}	MRAR
1→16	32.3	460.	209.	490.	242.	417.	29.	7.2
9→16	18.8	186.	164.	345.	132.	50.9	17.7	2.9
11→16	14.7	147.	128.	242.	107.	33.9	15.1	2.25
13→16	6.1	109.6	106.	196.	106.	32.3	5.6	5.8
14→16	1.2	47.5	105.9	123.9	62.9	24.8	0.92	27.

Table 3.6 Typical values of distances for reference speaker WM. The 16th order model. (N^1-1) - parameters excluded from computation

in the computer memory) is determined. If this ratio termed "Acceptance Ratio" is greater than 2, then the identity claim is rejected; on the contrary if the acceptance ratio is less than 2, the identity claim of test speaker is accepted. The decision making process is presented in the Figure 3.2.

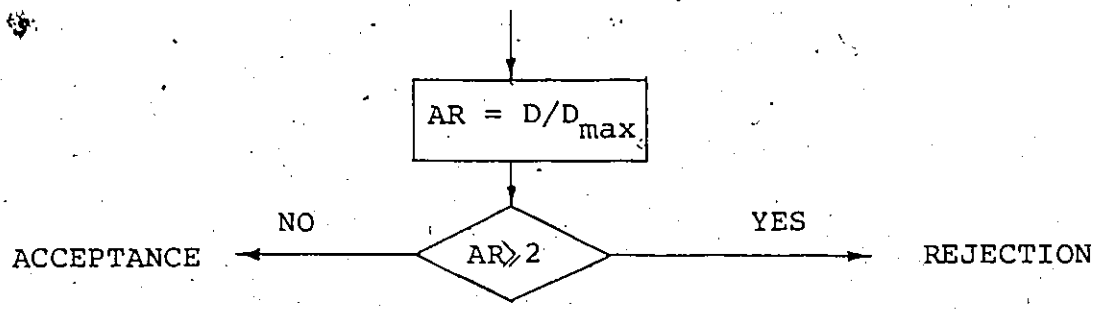


Fig. 3.2. Decision Process

3.9 RESULTS

Our text-dependent speaker verification system for high quality speech samples has been based on the sentence "We were away a year ago." spoken six times by each of six persons on six separate sessions equally distributed over one and a half months.

As it was mentioned in previous sections, the largest MRAR - coefficients appear when only the last orthogonal parameters are taken into consideration.

It confirms the observation of eigenvalues of orthogonal parameters that indicated very small statistical variance of higher order orthogonal parameters.

Hence, only these parameters should be used into distance computation as the most representative features.

The 8, 12 and 16th order of linear prediction model have been applied to analyze the prescribed sentence.

Every sentence was divided into 100 frames to provide appropriate time alignment.

The experimental results have shown, that using 8th order model and only last two orthogonal parameters the verification score reaches 100%.

The 12th order model gives the maximum verification accuracy when last three parameters are included.

The 16th order model uses last four orthogonal parameters and gets 100% accuracy.

In the table 3.7 the results of verification procedure using three models of linear prediction are presented.

	Model of Linear Prediction									
	M=8			M=12			M=16			
ORTPAR	1-8	5-8	7-8	1-12	7-12	10-12	1-16	9-16	11-16	13-16
Accuracy(%)	88.1	92.8	100	92.4	93.9	100.	91.3	92.2	95.2	100

Table 3.7 Verification score vs orthogonal parameters used for distance computation

The results obtained in this study indicate that very high accuracy of speaker verification system can be achieved using only the 8^{th} order model of inverse filter, which is the best choice from point of view of computational effort.

The verification experiment with just an 8^{th} order model was repeated to find out the effectiveness of proper time alignment.

Each digitized sentence has been divided into frames with fixed duration equal 20ms. Number of frames with fixed duration varied approximately between 60 to 100. The fixed frame duration 20ms was established to assume the stationarity of the speech process within this period.

Exactly the same verification procedure was applied to such prepared speech data. The accuracy of the system decreased dramatically to 86%. Practically every sixth comparison gave the wrong decision.

During the experiment it was observed that eliminating of silent portions and pauses within utterance slightly decreased the verification ability of the system.

Preliminary intuitive prediction that these pauses may supplementarily be informative about the person's speaking behaviour was firmly confirmed.

CHAPTER IV

EVALUATION OF SPEAKER VERIFICATION ALGORITHM IN PRESENCE OF NOISE

4.1 INTRODUCTION

An important problem in speaker verification arises, when speech inputs are noise corrupted. The versatile speaker verification system ought to be carried out on one of the existing telecommunication links to be less expensive, more convenient to use, applicable to real life requirements.

Practically, all communication channels transmitting the voice contain certain level of noise such as thermal noise, telephone exchange clipping, quantization noise, jamming of broadcasting, cross-talk distortions, interferences, etc. Signal-to-noise ratio (SNR) varies approximately from 30 to 10dB.

In this chapter the investigations are directed towards the evaluation of the sensitivity of orthogonal parameters to the level of noise in the speech signal.

Based on these results the accuracy of verification is determined, using only those parameters that are the least sensitive to noise.

Also, the influence of the order of the linear prediction model on verification accuracy will be studied.

4.2 NOISE GENERATION

Two methods of noise generation have been used in our experiment.

a) First, the infinite length of sequence of pseudo-random noise with zero mean i.e., DC level equal zero) within the bandwidth 0 ÷ 50kHz was produced by Hewlett-Packard type HP-3722 Noise Generator.

Since the speech signals, were bandpass filtered within 200÷3200Hz, and then sampled with 8kHz, it seemed to be natural to filter the noise sequence within the same band. Analog-to-digital conversion with 14 bits accuracy with sampling rate equal to 8kHz was then performed and the digitized, noise sequence of 5 sec. duration was stored in computer memory.

b) Second, the pseudo-random discrete noise was generated on NOVA-840 laboratory computer using slightly modified subroutines GAUSS and RANDU available in IBM-Scientific Subroutines Manual.

The mean value of noise sequence is assumed to be zero and variance corresponding to desired signal-to-noise ratio.

The noise samples with Gaussion distribution were added to the original speech samples to provide various controlled sequences of deteriorated speech.

The signal-to-noise ratio is defined in this work as

$$(4.2.1) \qquad S/N \text{ in dB} = 10 \log \frac{\sum_n s^2(n)}{\sum_n d^2(n)}$$

where $s(n)$ is the speech signal samples,
$d(n)$ is the additive noise,

and, the summation is over the the length of the whole utterance

It can be expressed in terms of variances of the signal and noise as

$$(4.2.2) \qquad S/N = 10 \log \frac{G_s^2}{G_d^2}$$

4.3 STUDY ON SENSITIVITY OF ORTHOGONAL PARAMETERS TO NOISE

It is already known, that linear prediction orthogonal parameters or at least part of them are extremely informative about speaker's identity and therefore, they are very useful tools in a practical speaker verification system for high quality speech inputs.

Now, the question arises, whether these parameters are adequate and applicable for noisy speech and how they will be altered by additive noise. To answer this question, the following procedure is proposed.

The high quality speech samples have been corrupted by the addition of an uncorrelated noise with zero mean and the variance corresponding to signal-to-noise ratio of 10, 15, 20 and 30dB.

Newly created noisy utterances were processed according to the procedure described in Chapter III.

First however, it seems to be interesting to analyze the noisy speech in order to obtain the parametric representation of a high quality speech.

4.3.1 Estimation of Linear Prediction Coefficients of a Clean Signal from Noisy Speech.

Let us denote as

$x(n)$ - clean speech signal

$\hat{x}(n)$ - noisy signal (with added noise)

where $\hat{x}(n) = x(n) + e(n)$

$e(n)$ - noise sequence (with desired variance Ge^2)

where $Ge^2 = E[e^2(n)] = E[(\hat{x}(n) - x(n)^2]$

$\hat{\hat{x}}(n)$ - predicted version of a clean signal derived from the noisy signal $\hat{x}(n)$

$$\hat{\hat{x}}(n) = \sum_{k=1}^{M} b_k \hat{x}(n-k) \tag{4.3.1}$$

$$d(n) = x(n) - \hat{\hat{x}}(n) \tag{4.3.2}$$

Now, we can express the variance of prediction error as

$$(4.3.3)\quad G_d^2 = E\left[d^2(n)\right] = E\left[(x(n)-\hat{x}(n)\right] =$$

$$= E\left[(x(n)-\sum_{k=1}^{M} b_k \hat{x}(n-k))^2\right] =$$

$$= E\left[(x(n) - \sum_{k=1}^{M} b_k x(n-k)^2 - \sum_{k=1}^{M} b_k e(n-k))^2\right]$$

To minimize the prediction error d(n) we set

$$\frac{\delta G_d^2}{\delta b_j}$$ to zero, obtaining M equations

$$(4.3.4)\quad \frac{\delta G d^2}{\delta b_k} = -2E\left\{\left[x(n)-\sum_{k=1}^{M} b_k\left(x(n-k)+e(n-k)\right)\right]\ \left[x(n-j)+e(n-j)\right]\right\}$$

$$= 0 \qquad \text{for } 1\leq j\leq M$$

or in a short form

$$(4.3.5)\quad E\left[(x(n)-\hat{x}(n))\,\hat{x}(n-j)\right] = E\left[d(n)\hat{x}(n-j)\right] = 0$$

Hence, d(n) and $\hat{x}(n-j)$ are uncorrelated for $1\leq j\leq M$.

Equation (4.3.4) can be written in the form of M equations

$$(4.3.6)\quad E\left[x(n-j)x(n)\right] + E\left[e(n-j)x(n)\right] =$$

$$\sum_{k=1}^{M} b_k E\left[x(n-j)x(n-k)\right]$$

$$+ \sum_{k-1}^{M} b_k E\left[e(n-j)x(n-k)\right]$$

$$+ \sum_{k=1}^{M} b_k E\left[x(n-j)e(n-k)\right]$$

$$+ \sum_{k=1}^{M} b_k E\left[e(n-j)e(n-k)\right]$$

Since e(n) and x(n) are statistically uncorrelated, the 2nd and 3rd components are equal to zero, and finally an expression in terms of autocorrelation function of a clean signal x(n) is obtained.

$$(4.3.7)\quad \hat{\phi}(j) = \sum_{k=1}^{M} b_k\left[\phi(j-k)+\delta(j-k)Ge^2\right]$$

for $1 \le j \le M$

In matrix form:

$$(4.3.8)\begin{bmatrix} \phi(0)+Ge^2 & \phi(1) & \cdots & \phi(M-1) \\ \phi(1) & \phi(0)+Ge^2 & \cdots & \phi(M-2) \\ \vdots & & & \\ \phi(M-1) & \phi(M-2) & \cdots & \phi(0)+Ge^2 \end{bmatrix} \begin{bmatrix} b_1 \\ b \\ \vdots \\ b_M \end{bmatrix} = \begin{bmatrix} \phi(1) \\ \phi(2) \\ \vdots \\ \phi(M) \end{bmatrix}$$

As one can see, the influence of additive noise altered only the diagonal of the matrix.

Linear prediction parameters a_k of the clean signal may be computed using the autocorrelation method, by solving a set of M-simultaneous equations written in matrix form.

$$(4.3.9)\quad \begin{bmatrix} \phi(0) & \phi(1) & \cdots & \phi(M-1) \\ \phi(1) & \phi(0) & \cdots & \phi(M-2) \\ \vdots & \ddots & \ddots & \vdots \\ \phi(M-1) & \phi(M-2) & \cdots & \phi(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_M \end{bmatrix} = \begin{bmatrix} \phi(1) \\ \phi(2) \\ \vdots \\ \phi(M) \end{bmatrix}$$

The left sides of both matrix equations are equal, hence,

$$(4.3.10)\quad \begin{bmatrix} \phi(0) & \phi(1) & \cdots & \phi(M-1) \\ \phi(1) & \phi(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \phi(1) \\ \phi(M-1) & \cdots & \phi(1) & \phi(0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \vdots \\ a(M) \end{bmatrix} = \begin{bmatrix} \phi(0)+Ge^2 & \cdots & \phi(M-1) \\ \phi(1) & \ddots & \vdots \\ \vdots & \ddots & \phi(1) \\ \phi(M-1) & \cdots & \phi(0)+Ge^2 \end{bmatrix} \begin{bmatrix} b(1) \\ b(2) \\ \vdots \\ b(M) \end{bmatrix}$$

It can be represented in the short form

$$(4.3.11)\quad \Phi \cdot A = \left(\Phi + Ge^2 \cdot I\right) \cdot B$$

where I - identity matrix

Let's define the autocorrelation matrix Φ_N for noisy signal as

$$\Phi_N = \Phi + Ge^2 \cdot I \tag{4.3.12}$$

then the equation (4.3.13) may be rewritten as

$$\left(\Phi_N - Ge^2 \cdot I\right) A = \Phi_N \cdot B \tag{4.3.13}$$

and we get important relations between a_k's and b_k's.

$$A = \left(\Phi_N - Ge^2 I\right)^{-1} \Phi_N \cdot B \tag{4.3.14a}$$

or

$$B = \Phi_N^{-1}\left(\Phi_N - Ge^2 I\right) \cdot A \tag{4.3.14b}$$

Now, assuming that the variance of noise is known, the processing of noisy data can make it possible to estimate a set of linear prediction coefficients equivalent to one obtained from high quality speech data.

There is one restriction: the existence of the inverse matrix $\left(\Phi_N - Ge^2 I\right)^{-1}$.

If it does not exist, the problem is unsolvable. In a practical system, since the analysis is performed in a frame by frame manner, the particular frame for which the mentioned inverse matrix doesn't exist is simply discarded from further computations.

4.3.2 Relation Between Orthogonal Parameters of High Quality Speech and Noise Added Speech

Let us define the orthogonal parameters of a clean signal s(n) computed over reference equivalents $V_{i_{ref}}$ derived from high quality speech.

$$(4.3.15) \quad \psi_s ij = \sum_{k=1}^{M} v_{ki_{ref}} \cdot a_{kj}$$

where

$$\left.\begin{matrix} i \\ k \end{matrix}\right\} = 1,2,\cdots \text{ M-order of prediction}$$

$$j = 1,2,\cdots \text{ L-number of frames}$$

and

$$(4.3.16) \quad \psi_z ij = \sum_{k=1}^{M} v_{ki_{ref}} \cdot b_{kj}$$

as orthogonal parameters of noisy signal computed through the same set of reference eigenvectors.

In the matrix form the vectors of the orthogonal parameters will be expressed by

$$(4.3.17) \quad \Psi_s = A^T V_{ref}$$

and

$$(4.3.18) \quad \Psi_z = B^T V_{ref}$$

The transposition of matrix A from eq. (4.3.14a) gives

$$A^T = B^T \Phi_N{}^T \left[\left(\Phi_N - Ge^2 I\right)^{-1}\right]^T \tag{4.3.19}$$

Since for symmetric matrices $\Phi_N = \Phi_N{}^T$, hence

$$\Psi_s = B^T \Phi_N \left[\left(\Phi_N - Ge^2 I\right)^{-1}\right]^T \cdot V_{ref} \tag{4.3.20}$$

As a result, we found the equation that expresses the orthogonal parameters of a clean voice signal in terms of parameters derived from noisy signal.

4.3.3 The Measure of Sensitivity of the Orthogonal Parameters to Additive Noise

The difference vector Δ with M-elements $\delta_1, \delta_2 \cdots \delta_M$ is expressed as

$$\Delta = \Psi_s - \Psi_z \tag{4.3.21}$$

and using our previous formulations for matrices Ψ_s and Ψ_z

$$\begin{aligned} \Delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_M \end{bmatrix} &= B^T \cdot \Phi_N \left[\left(\Phi_N - Ge^2 \cdot I\right)^{-1}\right]^T \cdot V_{ref} - B^T \cdot V_{ref} \\ &= B^T \left\{\Phi_N \left[\left(\Phi_N - Ge^2 \cdot I\right)^{-1}\right]^T - I\right\} \cdot V_{ref} \end{aligned} \tag{4.3.22}$$

Once the elements δ_i of a vector Δ are computed, the measure of the sensitivity of ORTPARs to additive noise can

be performed using a simple (percentage) error defined as

$$e_i = \frac{\delta_i}{p_{i_s}} 100[\%] \qquad \text{for } 1 \leq i \leq M \tag{4.3.23}$$

4.3.4. Experimental Results

For the purpose of the experiment, the high quality speech samples have been corrupted by the addition of pseudo-random wideband noise with zero mean value and the variance corresponding to SNR of 10, 15, 20 and 30dB.

The inverse filter coefficients calculation and the orthogonalization procedure followed by the computation of the orthogonal parameters have been performed for each case of SNR.

The sets of ORTPAR's were computed via reference eigenvectors derived from noise free speech in all cases.

The examples of the percentage variation in the orthogonal parameters for the 8^{th}, 12^{th} and 16^{th} order of linear prediction models are displayed in Tables 4.1, 4.2, 4.3, 4.4 and 4.5.

As one can observe, the least sensitive parameters are also the least significant orthogonal parameters for all tested models. For the 8^{th} order model, the last two ORTPARs are the most resistant to noise influence.

Speaker WM6 Reference Speaker JN6

SNR (dB)

Orthogonal Parameters	30	20	15	10
1	-3.8	-21.	-37.	-55.
2	8.2	36.	56.	71.
3	210.	1266	2111.	2709.
4	-11.	-.44	-110.	-203.
5	2.4	7.5	7.9	1.5
6	1.2	8.3	19.6	40.
7	.52	3.3	6.5	11.7
8	-.001	.56	2.1	6.

Table 4.1. The percentage variation of the ORTPARs for noisy utterances calculated with respect to reference ORTPARs derived from high quality speech. The 8th order model.

Speaker JN6 Reference Speaker WM6

SNR (dB)

Orthogonal Parameters	30	20	15	10
1	-4.5	-25.	-43.	-62.
2	8.1	34.	47.	55.
3	-1.2	-4.1	-11.	-22.
4	-16.	-75.	-129.	-170.
5	4.	19.	23.5	8.7
6	1.01	7.04	15.2	29.
7	.28	1.8	3.5	6.8
8	.07	-.19	-1.6	-5.6

Table 4.2. The percentage variation of the ORTPARs for noisy utterances calculated with respect to reference ORTPARs derived from high quality speech. The 8th order model.

Speaker WM6 Reference Speaker WM6

SNR (dB)

Orthogonal Parameters	30	20	15	10
1	-7.6	-36.3	-57.5	-76.5
2	15.6	-2.4	-87.	-208
3	-31.9	-154	-192	-166
4	13.5	54.3	74.7	88.8
5	-.41	22.3	55.7	78.
6	-50.7	-188	-241	-248
7	2.7	-.26	-11.3	-27.1
8	-4.0	-13.6	-21.7	-34.8
9	1.4	9.7	20.4	36.4
10	.33	.71	.62	4.3
11	.68	2.61	1.72	-6.4
12	-.19	-.79	-.72	.57

Table 4.3. The percentage variation of the ORTPARs for noisy speech with respect to reference ORTPARs derived from high quality speech, the 12th order model.

Speaker PA1 Reference Speaker WM6

SNR (dB)

Orthogonal Parameters	30	20	15	10
1	-14.7	-44.9	-64.2	-80.8
2	41.7	86.9	99.7	101
3	32	58.7	68.5	76.2
4	24.5	55.5	70.8	83.3
5	52.6	151	239	310
6	-39.4	-121.	-164.	-190
7	13	8.7	-3.9	-20.4
8	-.19	-3.3	-11.	-24.4
9	2.2	10.6	21.5	37.7
10	.30	.46	.79	4.9
11	-.11	.50	-1.4	-10.1
12	-.83	-1.9	-1.9	-.34

Table 4.4. The percentage variation of the ORTPARs for noisy speech with respect to reference ORTPARs derived from high quality speech. The 12th order model.

Speaker WM6 Reference Speaker WM6

SNR (dB)

Orthogonal Parameters	30	20	15	10
1	-7.7	-37	-58	-76
2	26.9	-7.4	-112	-277
3	-27.7	-133	-163	-139
4	14.9	69	98	116
5	-26.3	-79	-78	-61
6	-14.2	-59	-78	-77
7	-3.1	-23	-40	-56
8	.88	-5.7	-17	-33
9	-5.0	15	20	11.8
10	5.1	23	34	47
11	1.9	13.8	27	45
12	.3	4	10.5	22
13	.01	.98	2.4	8.8
14	.68	1.65	-1.21	-12
15	-.17	-.98	-.92	.12
16	.41	1.6	1.8	.61

Table 4.5. The percentage variation of the ORTPARs for noisy speech with respect to reference ORTPARs derived from high quality speech. The 16th order model

Speaker PA 1 Reference Speaker WM6

SNR(dB)

Orthogonal Parameters	30	20	15	10
1	-16.3	-53	-72	-131
2	50.3	78	123	245
3	43.2	62	103	185
4	28.7	39	87	173
5	21.3	31	65	127
6	-12.3	22	38	101
7	-7.7	-16	-45	-93
8	-5.8	-12	-34	-86
9	8.4	13	25	76
10	4.8	15	24	66
11	2.3	7	18	49
12	1.1	5	17	28
13	.97	3	11	13
14	.97	3.4	9	15
15	-.37	-0.8	-1.8	-2.4
16	-.25	-1.2	-1.9	-2.0

Table 4.6. The percentage variation in the ORTPARs for noisy speech.

In the case of 12th order model, last three and for 16th order, last four ORTPARs have demonstrated very small sensitivity to noise. The largest changes have occurred in the most significant ORTPARs which contain the major part of the linguistic content of the speech signal.

On the other hand, it is already known, that the least significant ORTPARs are the most indicative about speaker's identity. Thus the orthogonal parameters relevant to the speaker's identity are still preserved even when the utterances are corrupted by wideband noise.

It, therefore, seems possible that speaker verification algorithm will perform satisfactorily even when the sentences are noise corrupted.

4.4 VERIFICATION ANALYSIS

In this study, the speaker verification procedure was applied to noisy test utterances. In order to check whether the conclusions from the previous section are applicable, experiments have been done again with three orders of linear prediction model of speech and different levels of noise. Signal-to-noise ratio varies from 10÷30dB. The range 10÷15dB has been scanned every 1dB and the range above 15dB, every 5dB.

One of the goals was to establish the minimum SNR at which the accuracy of verification still remains the same as for high quality speech.

The results of the verification study are displayed in Table 4.7. In this table, the lowest signal-to-noise ratios at which the MRAR parameters exceeded 2 are shown for one reference speaker.

It is observed, that the best results are obtained when only the least significant orthogonal parameters are used in the distance computation.

Also the lowest SNR at which the accuracy deteriorated, is seen to depend on the order of the model used. This critical signal-to-noise ratio varied from 18dB for an 8^{th} order model to 12dB for 12^{th} and 16^{th} order models. In other words, the verification system accuracy has been still maintained until SNR reaches the above mentioned values. Since the use of the 16^{th} order model did not significantly improve the accuracy, it was concluded that a 12^{th} order model and the set of last three orthogonal parameters was the best choice from the point of view of computational effort and accuracy of verification.

The overall accuracy of verification obtained by this procedure, when all the speakers were included, was 100 percent.

4.5. SUMMARY

a) It has been shown, that the use of minimum reject-to-accept ratio (MRAR) in verification algorithms, simplifies the decision making process.

	ORDER of LINEAR PREDICTION									
	M-8			M=12			M=16			
Significant Orthogonal Parameters	1÷8	5÷8	7÷8	1÷12	9÷12	10÷12	1÷16	9÷16	11÷16	13÷16
MRAR	2.2	2.02	2.7	3.6	3.1	3.4	5.8	2.7	2.1	2.1
SNR(dB)	30	30	18	30	20	12	30	30	20	12

Table 4.7. Verification ability in function of the lowest possible SNR and number of significant ORTPARs

b) The sensitivity study on orthogonal parameters derived from noisy speech, has enabled the selection of the proper orthogonal parameters to be used in distance computation.

c) It has been demonstrated, for speech corrupted by wideband noise, that satisfactory verification can be realized even when SNR is as low as 12dB.

d) It has been found, that the 12th order model of linear prediction was the best choice from the point of view of verification accuracy and computational effort.

CHAPTER V
SPEAKER VERIFICATION SYSTEM OVER TELEPHONE LINES

5.1 INTRODUCTION

In view of the encouraging results obtained with speech inputs corrupted by additive wideband noise, it was decided to evaluate the accuracy of verification for speech inputs obtained from telephone channels.

In previous chapters, evaluations of the system have concentrated on investigating features to be analyzed and developing the comparison procedures to make the system as effective as possible in terms of increasing overall accuracy of verification.

The test utterances were recorded in quiet room using Hi-Fi microphone with flat frequency response over the range 40Hz÷16kHz.

However, there are several limitations of the quality of speech transmitted over telephone channel.

The telephone transmission is bandlimited with 200 to 3200Hz. Changing transmission conditions involves signal alteration over telephone line.

The frequency response of different channels varies slightly, partially due to attenuation characteristic of the line and repeaters, partially due to nonlinear behaviour

of the carbon transducer commonly used in telephone network. Hence, spectral and phase distortions, and time-varying properties are likely to be encountered.

Additionally, the uncontrolled and degraded environmental conditions involving acoustic background noise and disturbances emitted at the customer's terminal and picked-up by telephone microphones were common during the recording sessions.

The mentioned problems have made the speaker verification system for telephone inputs quite difficult.

The purpose of this study was to determine how well the speaker verification system would operate under these broadened, "real world" conditions.

5.2 NOISE SOURCES

There are several sources of noise, which alter and distort the original voice signals passing through a telephone channel. They can be divided into two basic groups:

a) the noise generated around the remote speaker's terminal (like typing machines, paper rustles, and other office noise effects which undesirably affect the quality of the telephone conversation.

b) the noise generated in the line during transmission of the signal.

Since in our experiment, all the speakers have called from relatively noise free places we shall not consider the first group of noise as really important.

However, the second group must be considered as a significant source of degradation of speech samples.

It can be split into:

a) background noise:

thermic noise of line repeaters, telephone exchange clipping, crosstalk distortions, atmospheric disturbances, etc.

b) nonlinear distortions:

nonlinearity of the carbon microphones of types HA1 and M1 widely used in the telephone network, and nonlinear behaviour of line amplifiers.

It is commonly accepted that the carbon microphones basically have a limited life because of aging carbon. The sensitivity and resistance of the transducer are then time-varying parameters.

Moreover, the frequency response of the carbon microphone varies with its angular position, which is illustrated in Fig. 5.1.

All this shows that since every telephone call is initialized on a random, temporarily available pair of telephone lines and performed under different conditions, the transmitted voice signal is altered in the different way each time.

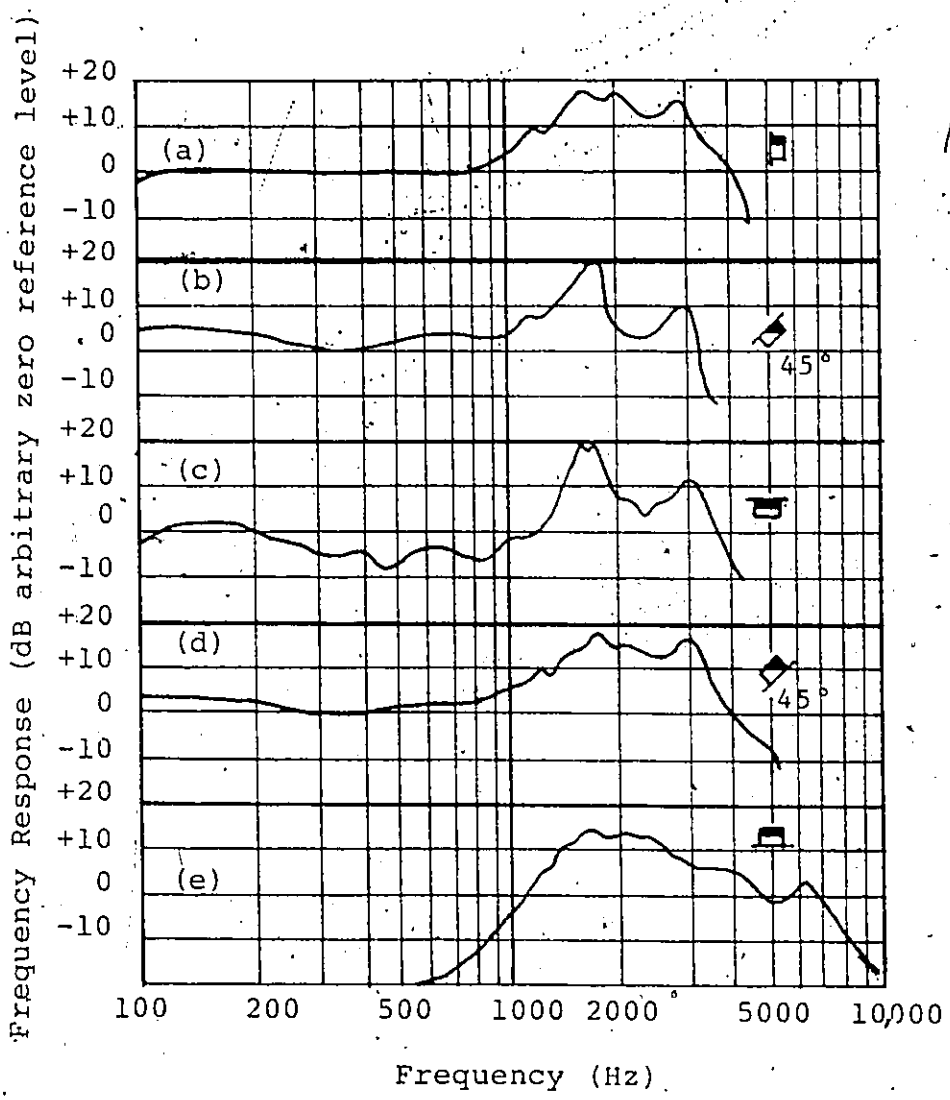


Fig. 5.1. Possible variation of carbon-microphone response with angular position.

It makes the analysis of the telephone speech quite difficult for speaker verification purpose.

5.3.1 Spectral Analysis of a Signal and Noise in the Telephone Line

For purpose of this analysis, two sessions were arranged to make parallel recordings of telephone and high quality speech simultaneously.

The question was, how the telephone media changes the original signal.

The high quality speech signal was taken as a pattern original signal.

The 2048 samples of the beginning of the sentence "We were away a year ago." sampled with the rate of 8kHz are shown in both versions on Fig. 5.2 and 5.3.

As one might observe, the distortions of the signal are much higher for telephone speech.

The granular noise, if occurs, increases with the amplitude of a signal. It means, that the generated noise is partially correlated with the signal.

5.3.2 Estimation of SNR

In order to measure the signal-to-noise ratio in the telephone line, two versions of the same speech utterance spoken simultaneously via telephone and microphone were compared.

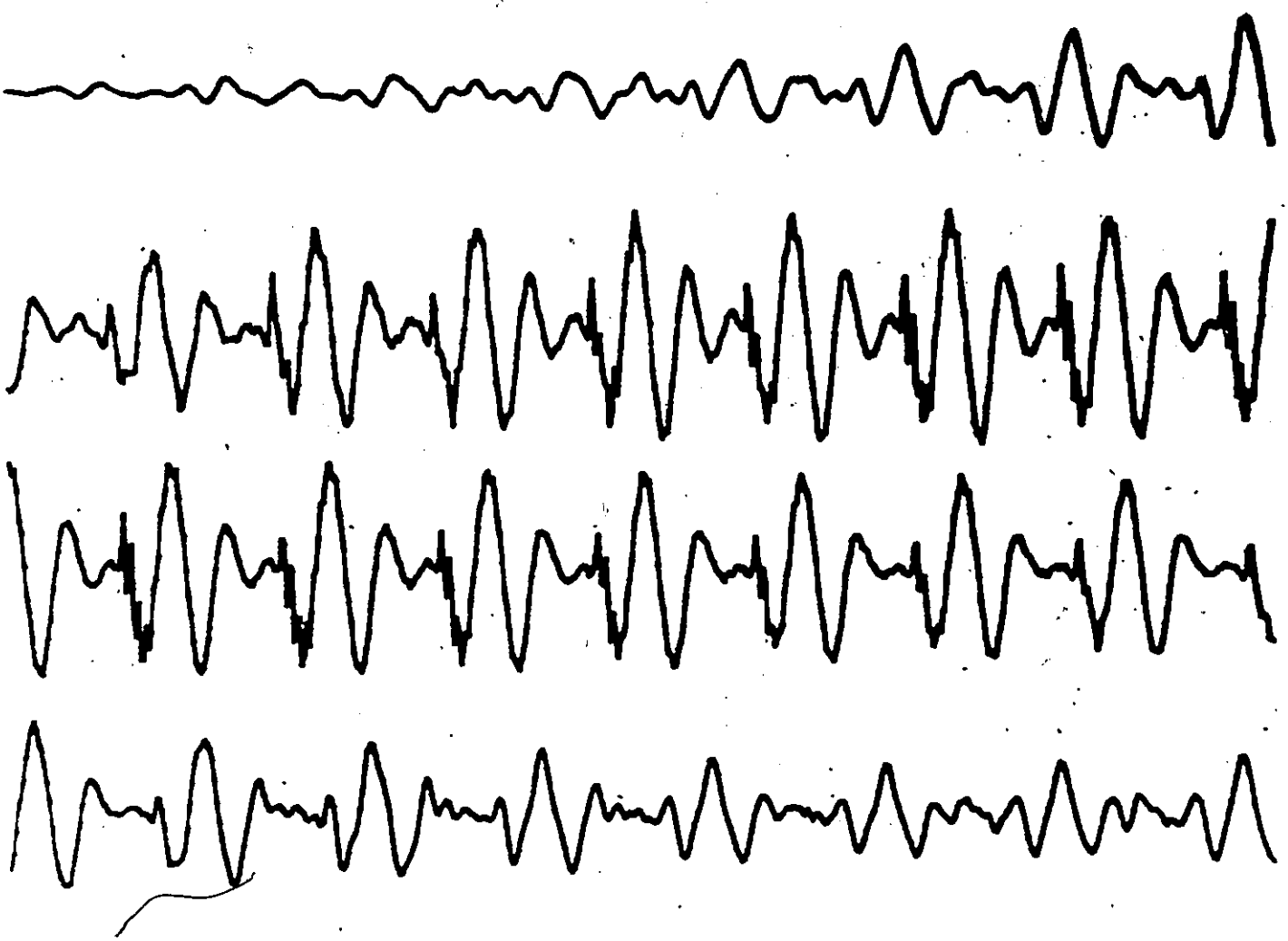


Fig. 5.2. The waveform of the 250ms period of the all-voiced sentence, "We were away a year ago." High quality speech.

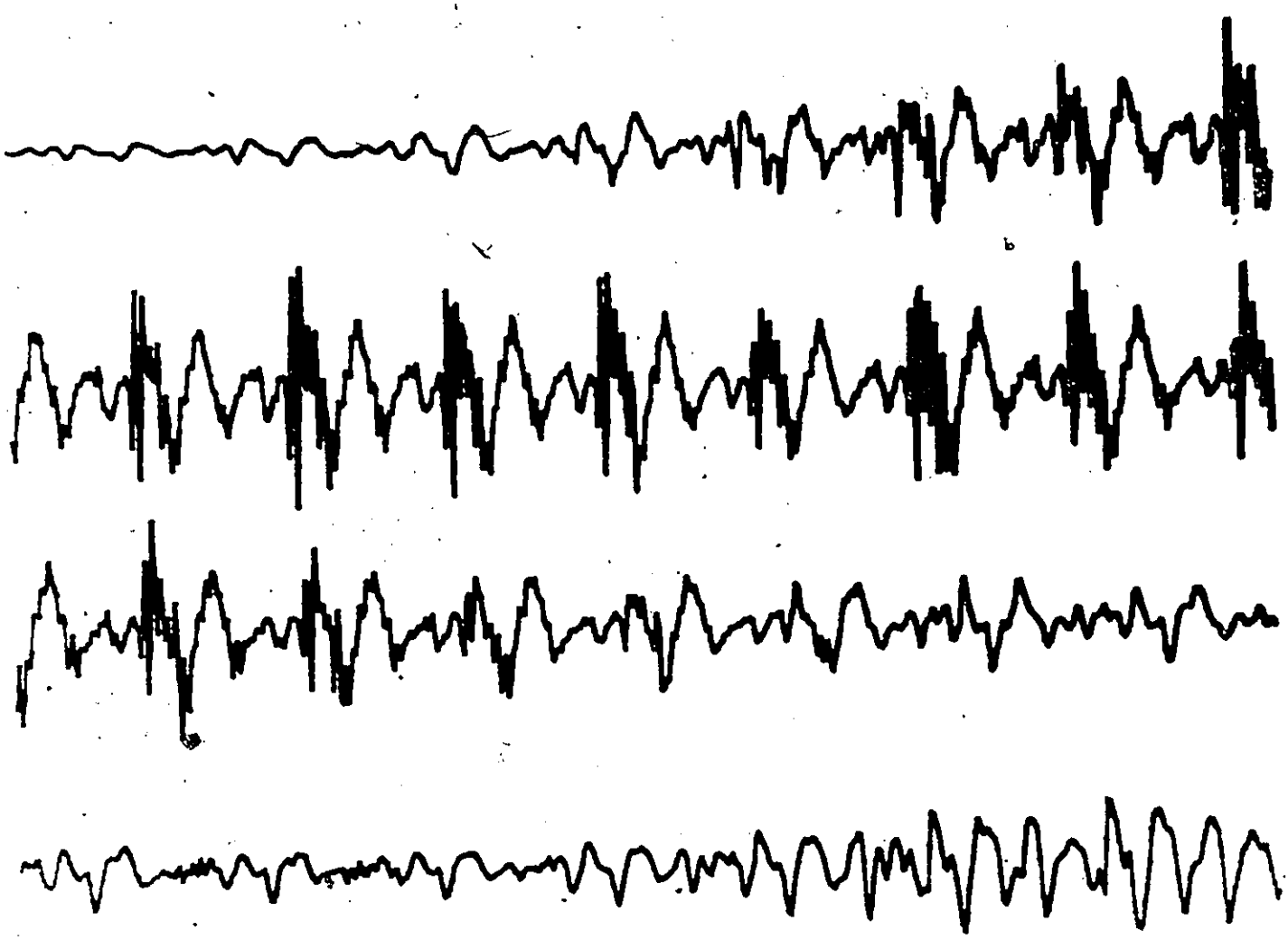


Fig. 5.3. The waveform of the 250ms period of the all-voiced sentence "We were away a year ago." Telephone speech.

The SNR in dB was defined as:

$$(5.4.1) \qquad SNR = 10\log \frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N} \left[x(n) - \alpha^{1/2} y(n)\right]^2}$$

where $x(n)$ - high quality speech

$y(n)$ - telephone speech

N - total number of speech samples in utterance

α - matching energy factor

The results obtained within the range 0÷3dB, were intuitively underestimated values.

It was mainly due to difficulties with a proper time alignment of both sections of analyzed speech, and phase distortions in telephone speech. In the light of impossibility of reliable estimation of SNR in time domain, it was decided to perform the computation in frequency domain.

Taking advantage of Parseval Theorem, that matches energy in both domains

$$(5.4.2) \qquad \sum_{n=-\infty}^{\infty} |x(n)|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{jw})|^2 \, dw$$

one can define SNR in frequency domain

$$(5.4.3)\quad SNR = 10\log \frac{\sum_{k=1}^{K/2+1} X^2(k)}{\sum_{k=1}^{K/2+1} \left[X(k) - \alpha^{1/2} Y(k)\right]^2}$$

where $X(k)$ - DFT of high quality speech

$Y(k)$ - DFT of telephone speech

α - matching energy factor

The Discrete Fourier Transform (DFT) of a finite duration time sequence $\{x(n)\}$ for $0 \leq M \leq N-1$, and its Inverse Transform are defined[39] as

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j(2\pi/N)nk} \qquad \text{for } k = 0,1,2\cdots,N-1$$

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j(2\pi/N)nk} \qquad \text{for } n = 0,1,2\cdots,N-1$$

The DFTs of high quality speech $X(k)$ and telephone speech $Y(k)$ were computed using Radix-2 Fast Fourier Transform (FFT) algorithm. The 256-point FFT scanned the desired band up to 8000Hz with the resolution of 31Hz, which was found sufficient for the purpose of this experiment. The computer program used is given in Appendix C. Since the FFT of the real data has the symmetry property, the only N/2+1 coefficients need to be used to signal-to-noise ratio computations.

The average SNR calculated using Eq. 5.4.3 was found to be around 16dB.

The average background noise during recording sessions was estimated as 21.5dB below the average signal level. The mean level of the calls initiated from the city was approximately 2-3dB below of those over from the local exchange.

The 33 percent of the telephone calls were originated from outside of the local university exchange.

5.3.3. Periodograms of a Signal and Noise

Considering the linear prediction technique in spectral analysis of a speech signal we deal with terms of the signal and model spectra [1].

The inverse filter A(z) gain constant G matches spectra energies of the model and the data using the autocorrelation method.

The speech signal spectrum $|X(e^{jw})|^2$ can be approximately expressed as the linear prediction model spectrum

$$\frac{G^2}{|A(e^{jw})|^2}$$

which is the smoothed version of the first one.

Hence,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}|X(e^{jw})|^2dw = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{G}{|A(e^{jw})|^2}dw \qquad (5.4.5)$$

The matching energy data factor G is equal to the total squared error α and is easily computed in the autocorrelation method. (See Appendix B)

The log magnitude spectrum of input data LM(X) and the log magnitude spectrum of the model LM(G/A) are formulated as

$$LM(X) = 10\log\left\{\left|X\left(e^{\frac{j2\pi k}{N}}\right)\right|^2\right\}$$

$$LM(G/A) = 10\log\left\{\frac{G^2}{\left|A\left(e^{\frac{j2\pi k}{N}}\right)\right|^2}\right\}$$

$$k=0,1,2\cdots\cdots N/2$$

In order to compute periodograms of the high quality and telephone signals both utterances were divided into a number of frames to give the duration of a frame N' less than N, equal to 256.

The input data from N'+1 up to N were filled by zeros.

Then, the 256-point FFT was applied in frame by frame manner and finally an average was computed over the whole sentence.

To compute the model spectrum estimation, the FFT of the sequence of linear prediction coefficients (a_0, $a_1, a_2 \ldots a_M$, 0,0, ...) appended with zeros up to 256 was applied.

The imaginary parts of input sequences were set naturally to zero.

The normalized periodogram provides the convenience of comparison, when the 0dB corresponds to the peak value of a "spectrum."

The examples of normalized periodograms of the high quality and telephone speech are shown in the Fig. 5.4 and Fig. 5.5, respectively. The FFT of the first 256 samples of the phrase "We were away a year ago." were computed and then log magnitudes versus frequency were plotted.

Looking at these graphs, one can say very little about the energy distribution and formants position over the band of 4kHz. Figures 5.6 and 5.7 show the log magnitude of a linear prediction models of the same segment of a high quality and telephone speech.

Here, the peaks representing formants are clearly seen on both periodograms.

The second formant F_2 for the telephone speech is only -3dB below first one, comparing to -13dB for original speech.

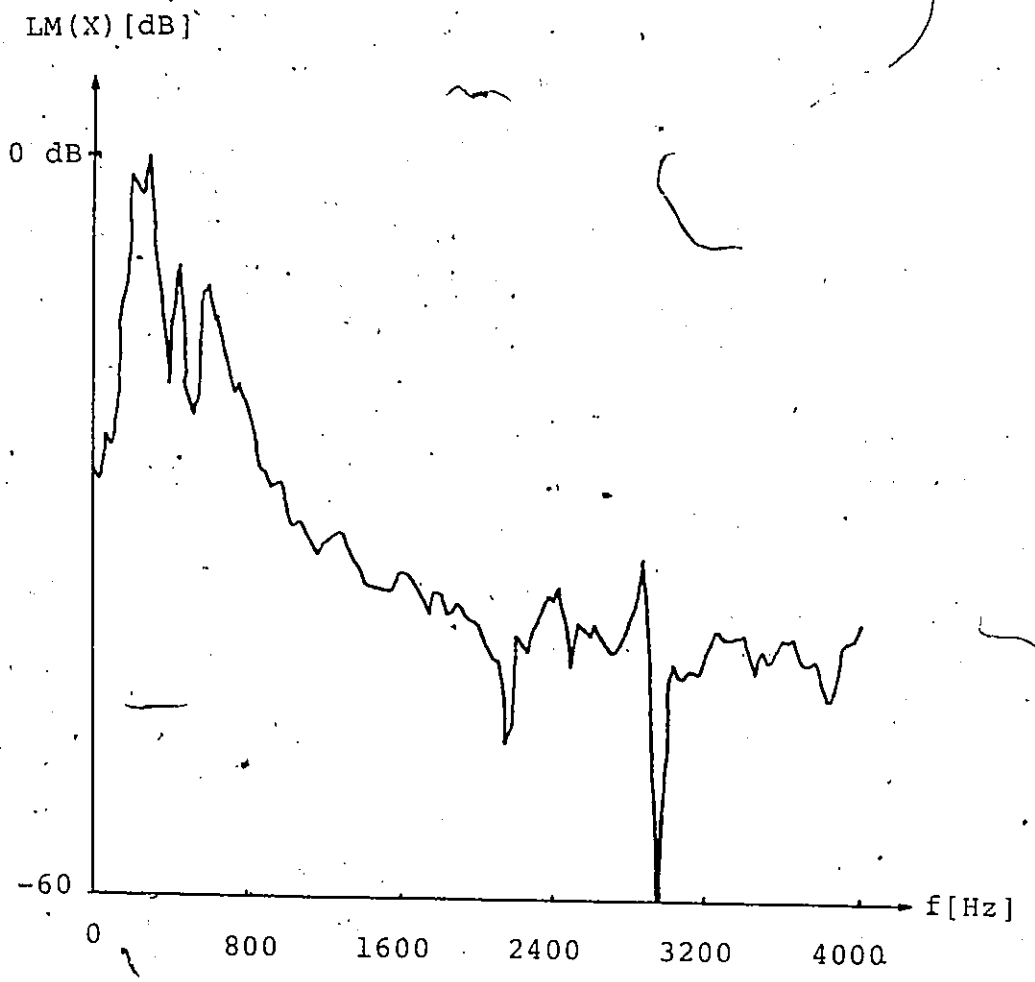


Fig. 5.4. Periodogram of first 256 samples of the word "We". High quality speech.

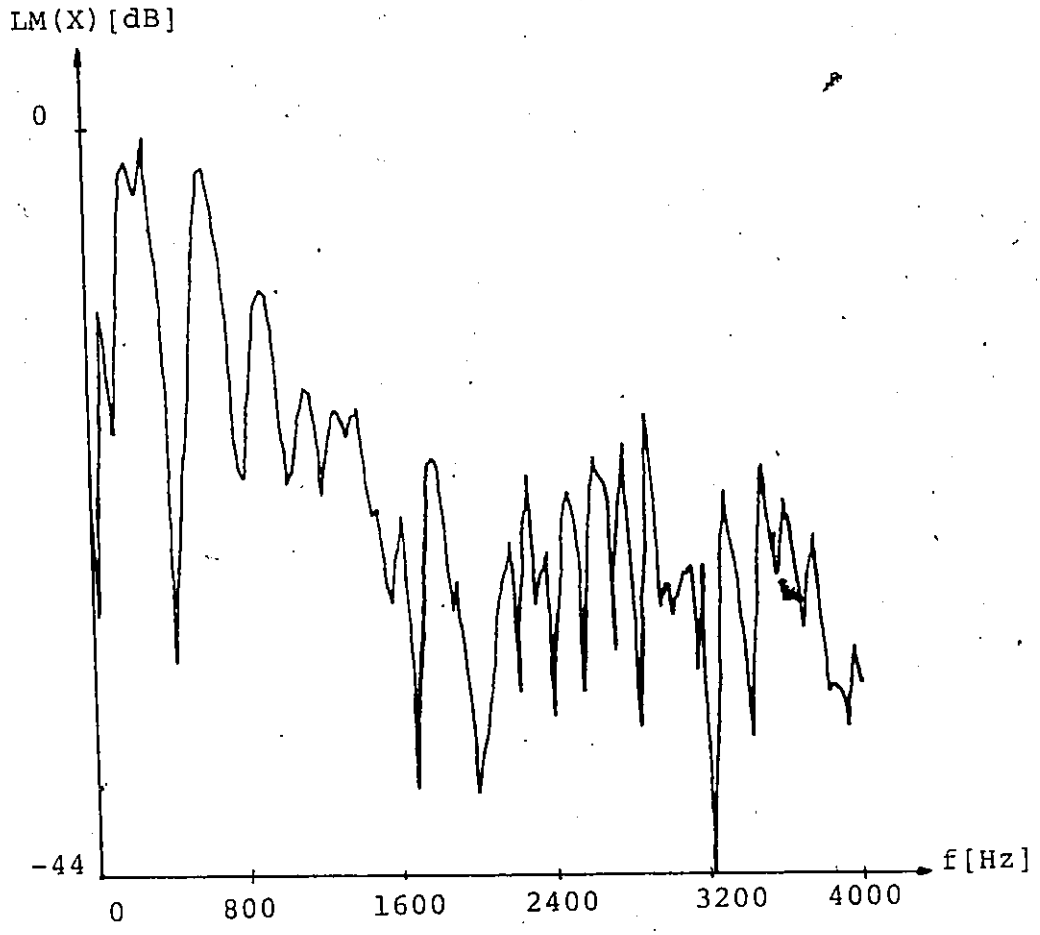


Fig. 5.5. Periodogram of first 256 samples of the word "We". Telephone speech.

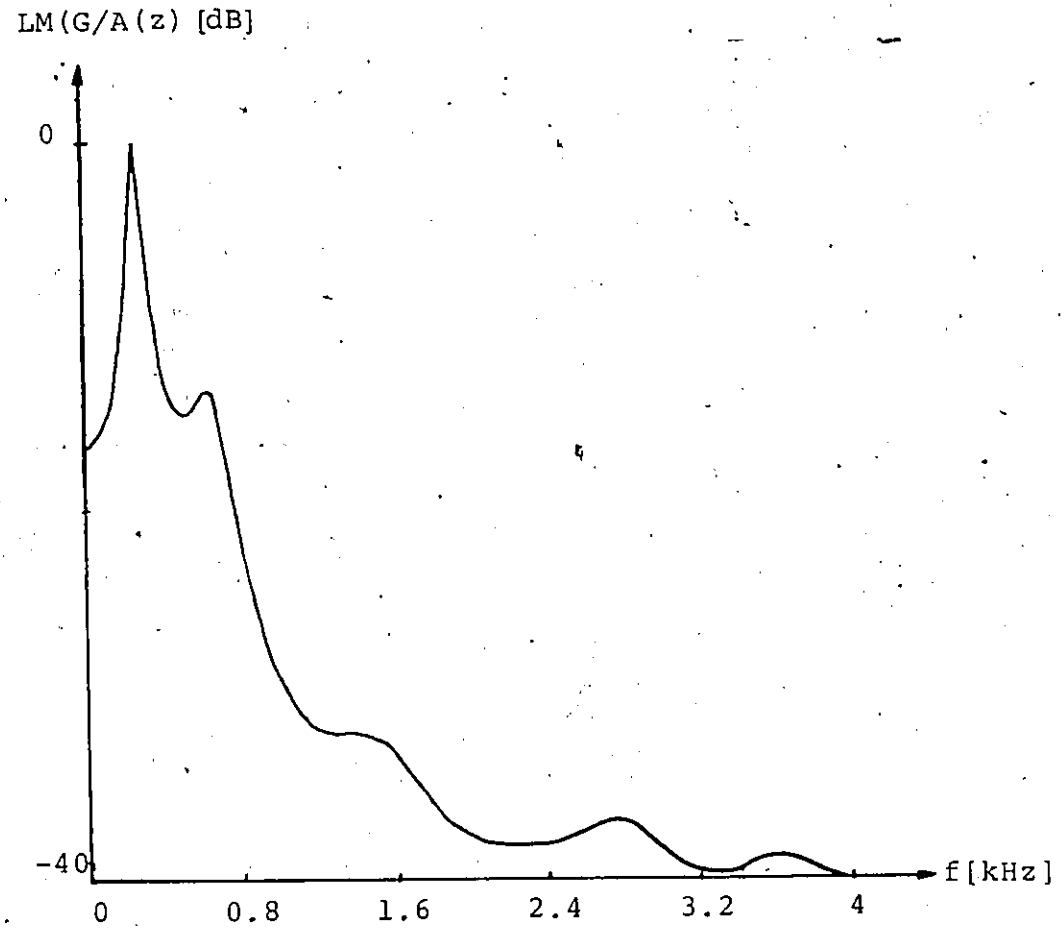


Fig. 5.6. Periodogram of spectral model G/A(z) of the word "We". High quality speech.

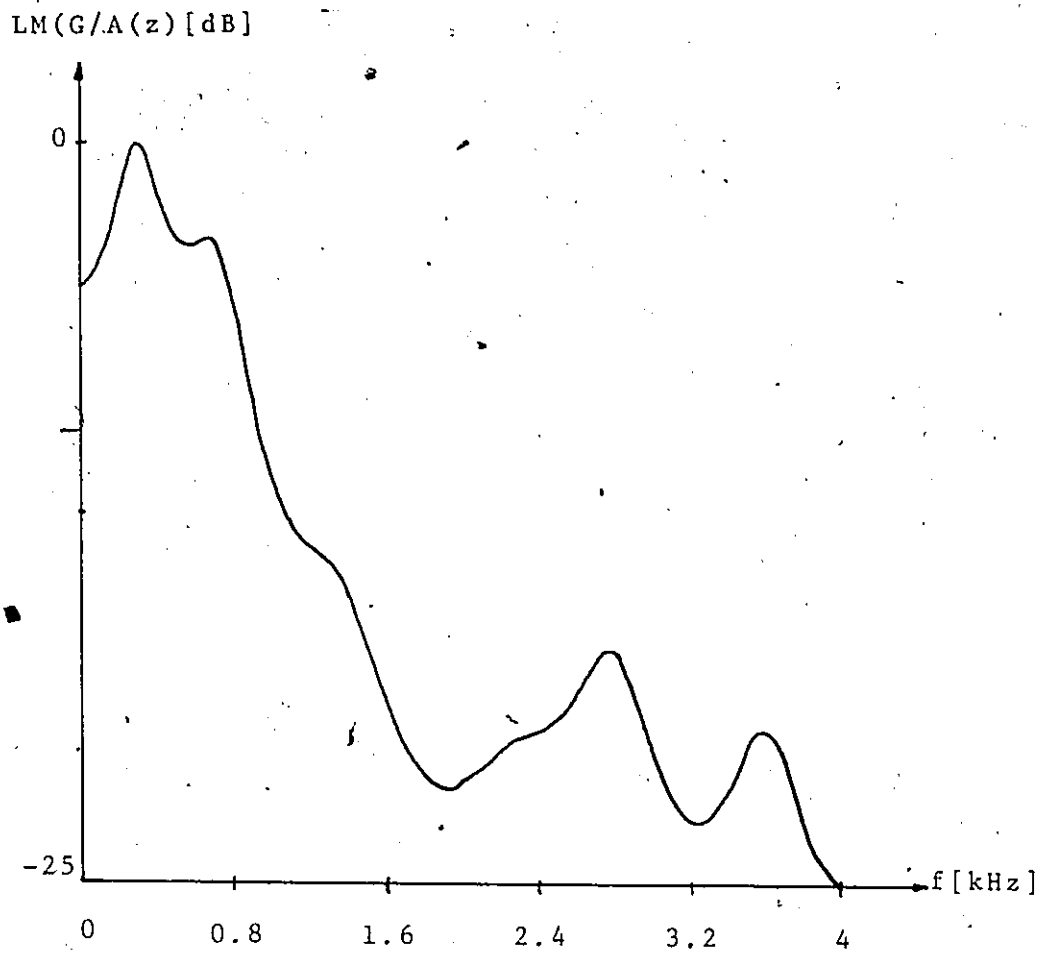


Fig. 5.7. Periodogram of spectral model G/A(z) of the word "We". Telephone speech.

Similarly, the third formant F_3 is -18dB down, versus -32dB for high quality speech. The log magnitude of total utterance of duration of approximately 16,000 samples, shown on Fig. 5.8 and 5.9 have indicated similar behaviour.

Results demonstrated, that the telephone channel characteristics have the general tendency to dump up high frequencies!

The channel acts as a kind of high pass filter but unfortunately with no fixed cut-off frequencies (see Chapter 5.3, Fig. 5.1).

On the Fig. 5.10 the periodogram of a noise in the idle telephone channel is presented. The 12,500 samples of a noise which corresponds to 1.5 sec. of duration were scanned. Dynamic range of the noise reaches 3.5dB.

5.4 SPEAKER VERIFICATION SYSTEM ACCURACY IN FUNCTION OF ORTHOGONAL PARAMETERS

The typical example of the behavior of MRAR coefficients versus number of orthogonal parameters used into distance computation for one reference speaker is presented in Table 5.1.

It was observed during the experiment that MRAR has usually been below CR=2, except one case only; when all orthogonal parameters were taken into consideration. It means, that the best results of verification are obtained

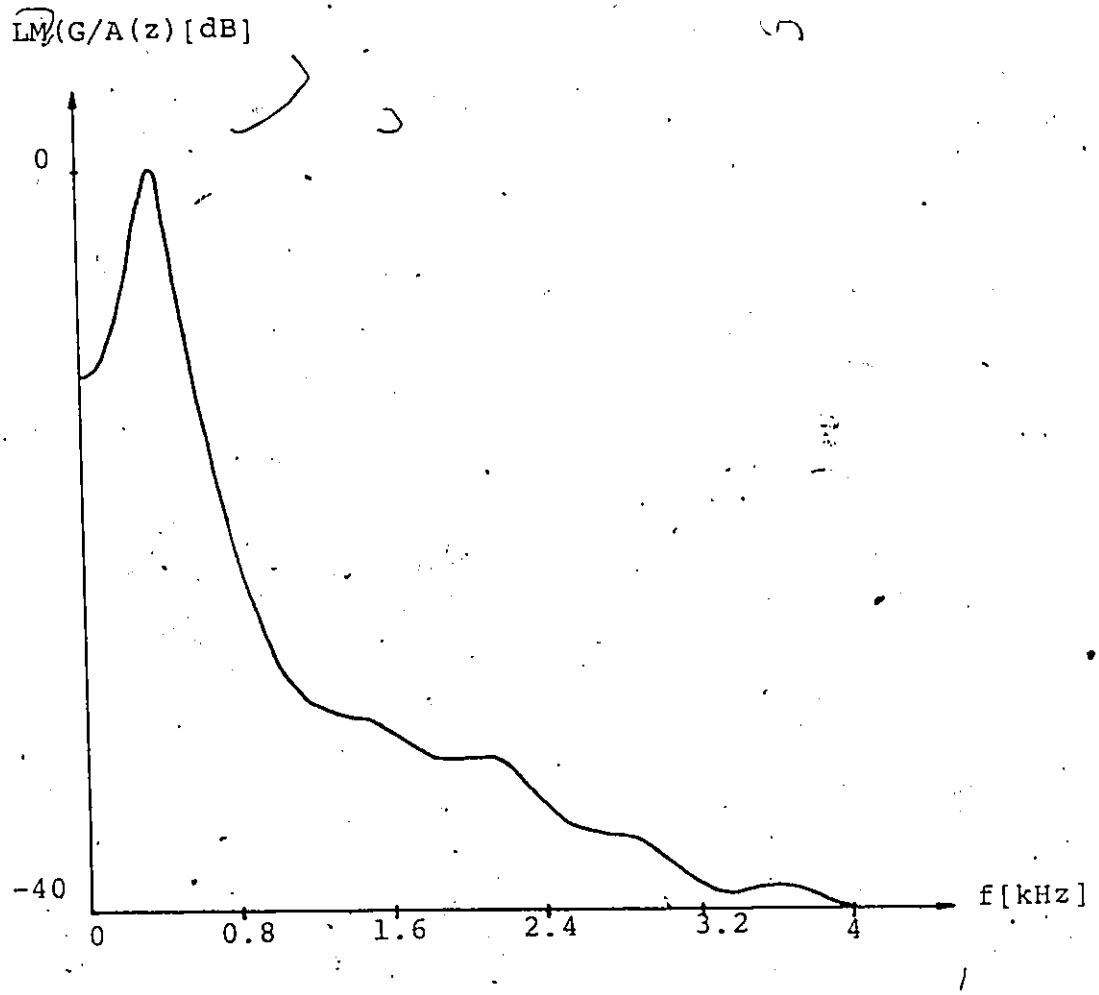


Fig. 5.8. Periodogram of spectral model G/A(z) of the sentence "We were away a year ago". High-quality speech.

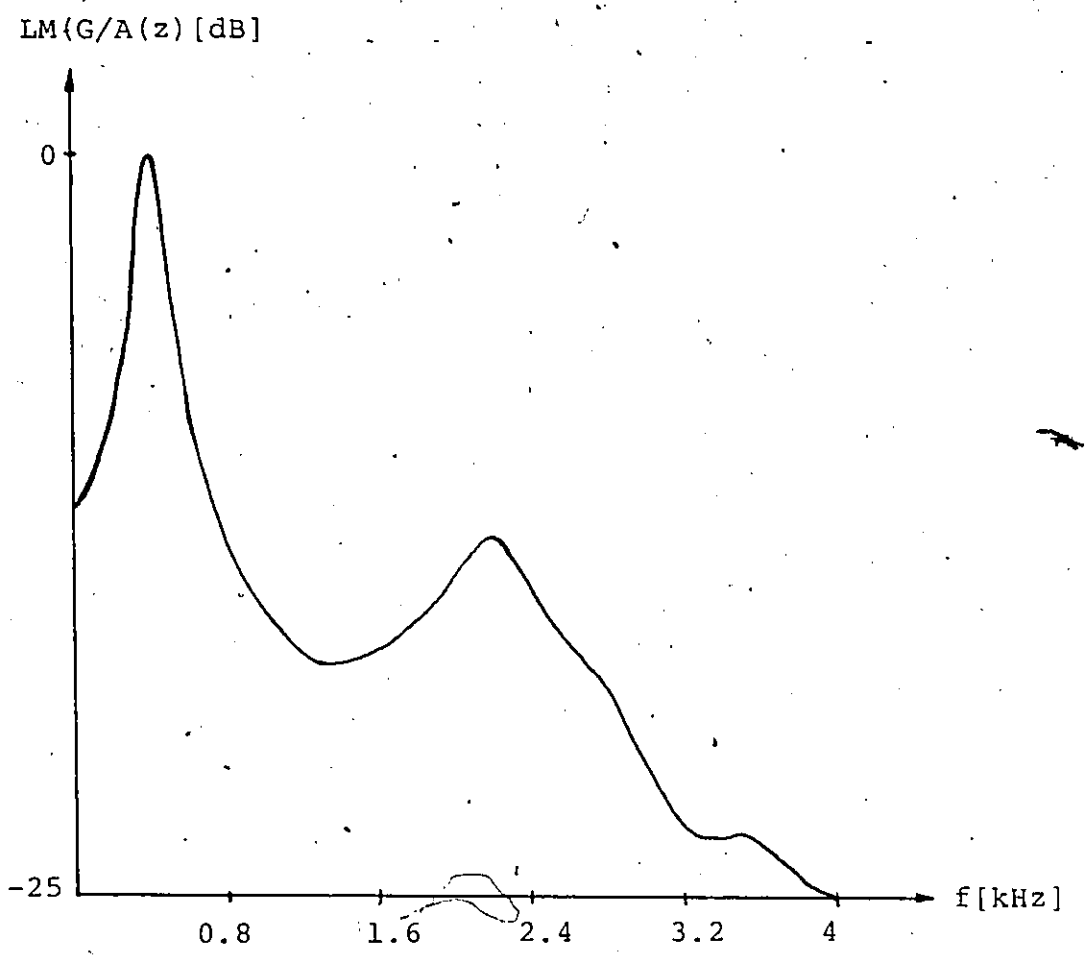


Fig. 5.9. Periodogram of spectral model G/A(z) of the sentence "We were away a year ago". Telephone speech.

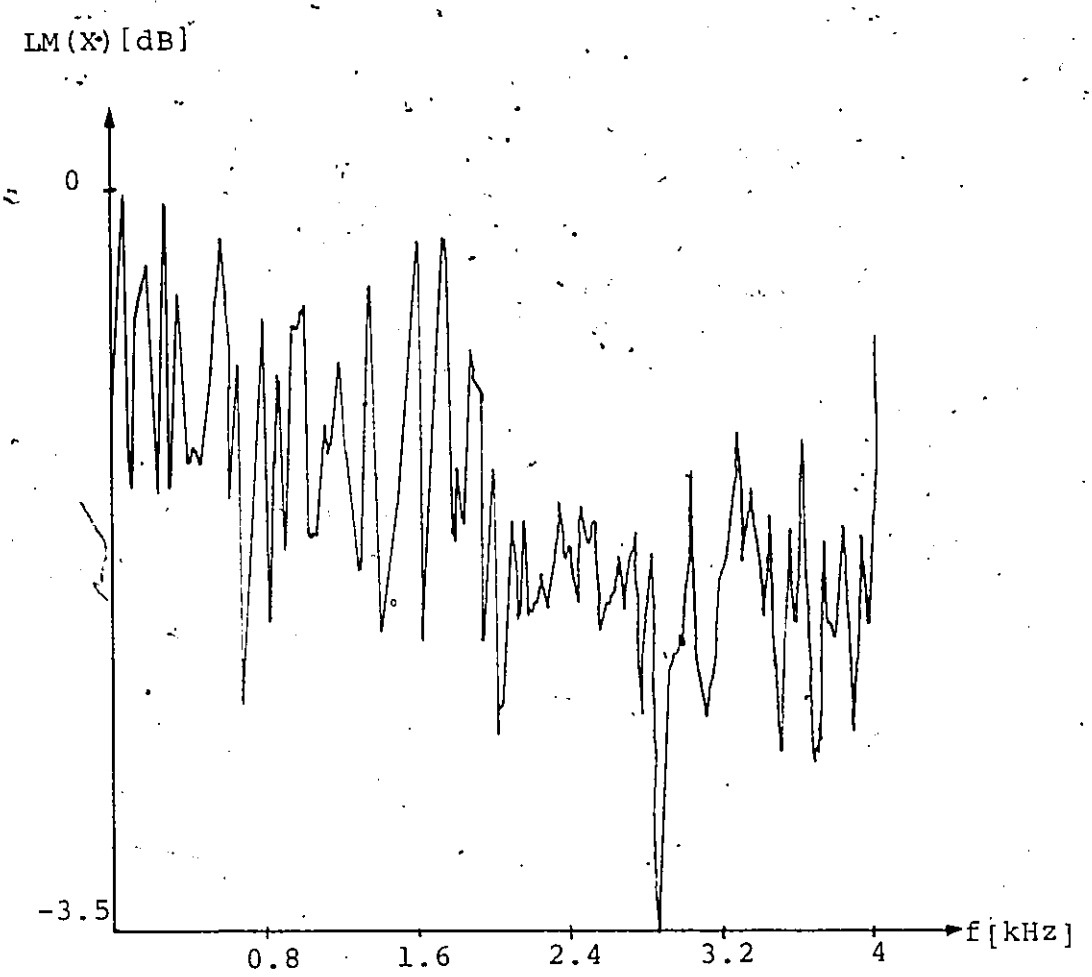


Fig. 5.10. Periodogram of the noise in the idle telephone channel.

when all 12 orthogonal parameters are used to distance measurement.

This behaviour is in contradiction with the results obtained for speaker verification for speech utterances corrupted by wideband random noise.

Some reasons for this difference are that:

a) the noise in telephone channel is not "white".

b) the noise power is distributed non-uniformly in the telephone bandwidth.

c) the noise is partially signal dependent

d) the reference orthogonal parameters were also derived from the telephone utterances.

	Number of used Orthogonal Parameters				
	1-12	5-12	7-12	9-12	10-12
MRAR	2.45	1.4	1.4	1.9	1.9

Table 5.1. The example values of MRAR coefficients versus the number of orthogonal parameters used in distance measurements.

Table 5.2 presents the results of the entire study of speaker verification accuracy as a function of the number of orthogonal parameters used. It can be seen that the highest accuracy (higher than 99%) is obtained only when all the orthogonal parameters are used.

Number of Orthogonal Parameters	Accuracy [%]
1	75.5
2	88.1
3	92.1
4	94.6
5	96.0
6	96.6
7	98.2
8	98.6
9	98.6
10	98.8
11	99.4
12	99.4

Table 5.2. Speaker verification accuracy for telephone inputs versus orthogonal parameters

Sentence number	Accuracy (%)
1	99.4
2	99.3
3	97.1
4	97.6
5	98.3
6	98.8

Table 5.3. Verification Accuracy over Telephone Lines. The 12th order of linear prediction--all orthogonal parameters included into distance computations.

Table 5.3 summarizes the result of the speaker verification procedure over telephone lines. The results show that the highest accuracy was obtained when all-voiced utterances (1 and 2) were used to discriminate among speakers. The sentences 3 and 4, containing plosive fricatives "p", "t", and significant amount of unvoiced sounds represent less the verification potential. Some reasons for these results can be, as it will be discussed in Appendix A, the simplified all-pole linear prediction model is a natural representation of non-nasal voiced sounds, but for nasals, and unvoiced fricatives,

the detailed acoustic theory requires both poles and zeros in the vocal tract transfer function.

5.5 VERIFICATION TIME

Duration of the test phrases varied from 1.5 to 2.5 seconds.

It is certainly interesting to know how long the total verification procedure will take. It will be shown in terms of duration of the test utterance as a time unit T. The processing time for each operation performed during the single verification is presented in Table 5.4.

Assuming that the mean duration of a test phrase is 2 sec., the whole verification can be performed within approximately 82 seconds.

As one may notice, the solution of M-simultaneous equations to get a set of linear prediction coefficients is the most time consuming operation.

The subroutine written in assembly language and the use of the array processor would significantly improve the efficiency of computations.

5.6 SUMMARY

An attempt was made to achieve close to the "real world" speaker verification system.

Processing Time For One Verification	
LPC extraction	22T
Eigenvector analysis	4T
Orthogonal parameters computation	3T
Comparison and Decision making	2T
Miscellaneous	10T (approx.)
TOTAL	41T

Table 5.4. Processing Time for one verification

Based on reference patterns derived from the telephone speech utterances, the system has been able to verify the person calling through the telephone with the probability of error less than 0.01.

In another evaluation of a speaker verification system over telephone lines, performed by Rosenberg [11], based on the pitch, intensity, the first three formants and selected linear predictor coefficients, the probability of an error 0.05 for adapted customers and approximately 0.1 for newcomers was achieved.

The memory requirements for one reference set for one speaker consist of 12 reference eigenvalues, 12 reference orthogonal parameters and matrix 12x12 of reference eigenvectors. All together -- 168 memory locations.

In possible future hardware implementation the use of PROM is suggested to make possible the periodic updating of the reference data (e.g. every few months) to reflect long-term variations in speaking behaviour of customers.

CHAPTER VI

SPEAKER VERIFICATION SYSTEM FOR SPEECH WITH UNKNOWN NOISE STATISTICS

6.1 INTRODUCTION

In most SVS algorithms the speaker dependent features for either reference or test speakers are extracted from speech signals uttered under similar ambient conditions. Typically the speech utterances for both the reference and test speakers were recorded in

1) an anechoic chamber
2) a relatively noise free environment (room conditions)
3) the same noisy environment
4) over telephone lines

Speaker Verification System operating with different speech inputs would require different sets of reference parameters.

However, in many practical situations, the reference speech samples are obtained under noise free conditions and it is often required to verify the identity of a speaker from an analysis of the person's speech uttered in a fairly noisy environment.

An application of some of the currently known verification algorithms has resulted in a high failure rate in terms of false acceptance and false rejection, thus restricting

the practical utility of these algorithms.

Our investigations have been directed towards finding such a speech preprocessing method prior to the extraction of speaker recognition features, which would enable us to get high accuracy of SVS.

6.2 VERIFICATION WITHOUT PRE-PROCESSING

Since the purpose of this study was to investigate the feasibility of verification algorithms when data are obtained under different ambient conditions, the following rules were applied to speech data preparation:

1) for the extraction of reference speaker dependent features high quality noise free speech recordings were utilized.
2) for verification purposes, test speaker's utterances were obtained from noisy telephone lines.

All speech utterances (high quality and telephone speech) were then filtered to conform to the telephone bandwidth of 200-3200Hz.

The filtered speech was then digitized to 14-bit resolution with sampling rate 8kHz and stored in the memory of NOVA 840 computer.

The verification algorithm described in Chapter III was used to evaluate the overall accuracy of verification. A 12^{th} order of linear prediction model was used.

The orthogonal parameters for test utterances were computed via the reference eigenvectors derived from high quality speech for a given speaker.

The results of a verification procedure are shown in the Table 6.1.

As one can see, the best verification accuracy obtained was 45.3%, when all the orthogonal parameters were included in distance computation.

This was clearly unsatisfactory and no improvements could be reached by changing the order of linear prediction model or the number of orthogonal parameters used in distance computations.

6.3 SPEECH PREPROCESSING

As it is known [13], the linear prediction analysis and synthesis performed on noisy speech changes the underlying structure of poles resulting in incorrect estimation of formants and their bandwidths of the clean speech signal.

The effect of addition of a wideband random noise results in a loss of resolution and the smoothing of a spectral estimate. This smoothing effect and displacement of the estimated poles are due to the introduction of spectral zeros by adding the noise to the clean speech [15].

Number of Orthogonal Parameters	Accuracy [%]
1	32.8
2	33.7
3	40.6
4	41.4
5	42.5
6	44.1
7	44.8
8	45.1
9	45.1
10	45.3
11	45.3
12	45.3

Table 6.1. Verification Accuracy of SVS from Telephone Speech Versus High Quality References.

Our experiments on the speech with additive wide-band pseudo-random noise confirmed these observations. The speech with SNR equal 0, 10 and 20dB was processed.

The sets of LPC parameters obtained from noisy and clean speech signals differ significantly.

The comparison of orthogonal parameters derived as a linear combination of LPC parameters for both noisy and clean speech did not give promising results as was noted in the previous section.

6.3.1 Enhancement of Speech Degraded by Noise

There are several techniques that have been proposed for enhancement of speech degraded by additive noise. They can be classified in three basic groups:

Technique 1 -- requires a priori knowledge about statistics of a noise. The use of Wiener filtering, when the power spectrum estimate of the noise is known (e.g., single tone interference, narrow-band background noise).

Technique 2 -- PNC (predictive noise cancellation) estimates the present background noise which is adaptively updated from an average all-pole noise spectrum. The spectrum is computed by averaging autocorrelations during non-speech activity [19].

Technique 3 -- ANC (adaptive noise cancellation) requires no information about statistics of a noise [16, 17].

When the spectrum of the signal under consideration changes in an unpredictable fashion, then adaptive filtering should be considered. Since the telephone speech contains the noise with time variant and unknown statistics then the adaptive noise cancelling method would be the most applicable to use.

6.3.2 Adaptive Noise Cancelling

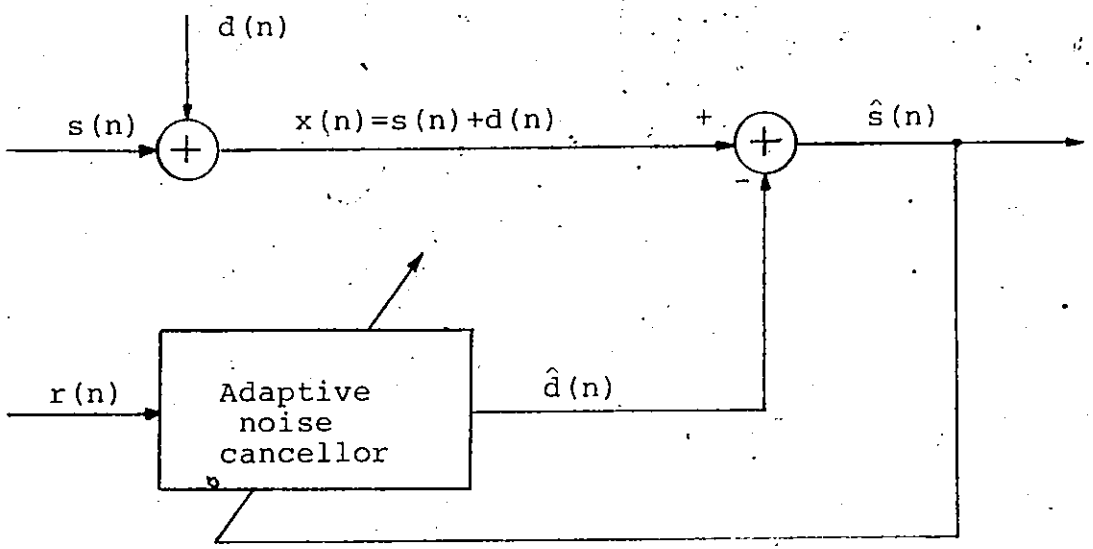
In our experiment, a self-tuning filter was used to eliminate the noise and improve the linear prediction analysis of the noisy speech signal [17].

The reference input signal $r(n)$ applied to the filter generates the estimate of the additive noise $d(n)$. This noise subtracted from the noisy signal $x(n)$ produces the system output signal $\hat{s}(n)$ which is the estimate of the clean signal $s(n)$.

The adaptive filter is self controlled by $\hat{s}(n)$ and tuned to give the minimum output system energy $E[\hat{s}^2(n)]$.

Let $$E\left[\hat{s}^2(n)\right] = E\left\{\left[s^2(n)\right] + \left[d(n)-\hat{d}(n)\right]^2 + 2s(n)\left[d(n)-\hat{d}(n)\right]\right\}$$

Since, it was assumed that signal $s(n)$ is uncorrelated to noise signals



x(n) - noisy signal

s(n) - clean signal

$\hat{s}(n)$ - noise cancelled signal

d(n) - background noise

$\hat{d}(n)$ - estimated noise

r(n) - reference input noise

Fig. 6.1. ADAPTIVE NOISE CANCELLING SYSTEM

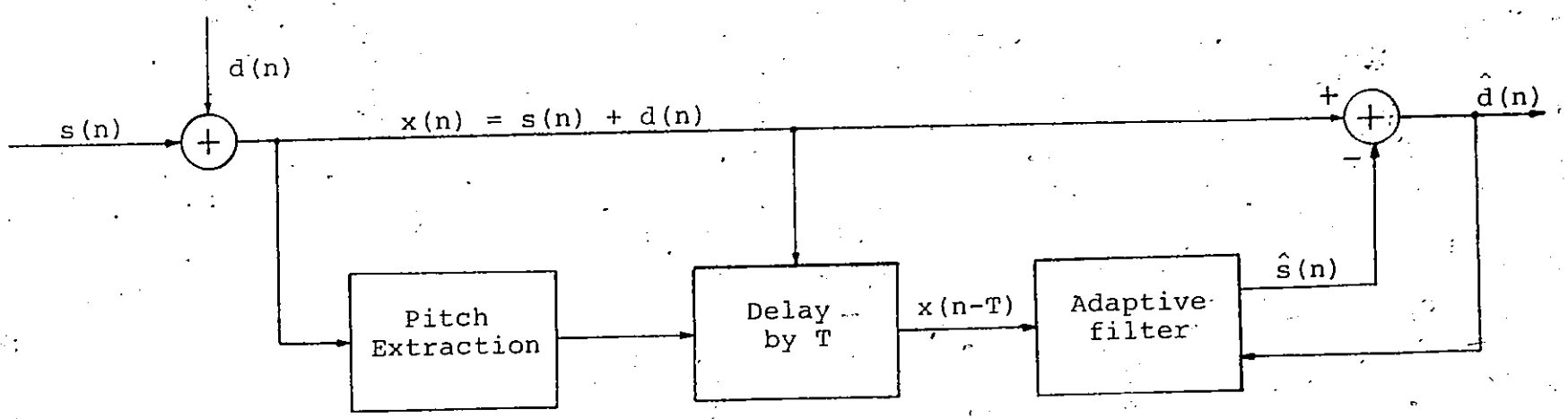


Fig. 6.2. MODIFIED ADAPTIVE NOISE CANCELLING SYSTEM

$$E\left\{s(n)\left[d(n)-\hat{d}(n)\right]\right\} = 0$$

and since $E\left[\hat{s}^2(n)\right]$ = const

hence,

minimum output energy is equal to

$$\min E\left[\hat{s}^2(n)\right] = \text{const.} + \min E\left[\left(d(n)-\hat{d}(n)\right)^2\right]$$

Therefore, if the output energy is minimized, then the noise $\hat{d}(n)$ is a best least squares fit to the true noise signal $d(n)$.

But since $\hat{s}(n)-s(n) = d(n)-\hat{d}(n)$

hence $\hat{s}(n)$ is a best least squares fit to the clean signal $s(n)$.

The generation of a reference input noise signal is a difficult problem, especially when one deals with a noisy telephone speech, where a noise is successively accumulated over transmission line.

However, taking advantage of the quasi-periodic nature of the speech, the signal $s(n)$ is highly correlated with delayed by pitch period T signal $s(n-T)$ and with advanced by T signal $s(n+T)$.

Moreover, the noise $d(n)$ is uncorrelated with noise $d(n\pm T)$ and signal $s(n)$.

The block diagram of noise cancellor is shown in Fig. 6.2.

Now, the delayed version of a noisy signal $x(n-T)$ is applied to the input of an adaptive filter. The output energy of a system is

$$E\left[\hat{d}^2(n)\right] = E\left[\left(x(n)-\hat{s}(n)\right)^2\right] = E\left[\left(s(n)+d(n)-\hat{s}(n)\right)^2\right] =$$
$$= E\left[\left(s(n)-\hat{s}(n)\right)^2\right] + 2E\left[d(n)\left(s(n)-\hat{s}(n)\right)\right] + E\left[d^2(n)\right]$$

Since the second component in above equation is equal zero and last one is a constant, the minimizing of the energy of $\hat{d}(n)$ results in a signal $\hat{s}(n)$, which is a best least-squares fit to the clean signal $s(n)$.

Self tuning noise canceller is non-recursive filter with the estimated clean signal $\hat{s}(n)$ as an output.

$$\hat{s}(n) = \sum_{i=0}^{M} b_i x(n-i-T)$$

for voiced frames.

For frames classified as unvoiced ones, the output $\hat{s}(n)$ was set equal to the original noisy signal $x(n)$.

Filter coefficients $\{b_i\}$ were updated according to least mean square (LMS) algorithm [16].

$$B_{n+1} = B_n + 2\mu\hat{d}(n)\chi_{n-T}$$

where B_n - set of coefficients $\{b_0, b_1, \ldots b_M\}$ at time n

χ_{n-T} - sequence $\{x(n-T), x(n-T-1), \ldots x(n-T-M)\}$

$\hat{d}(n) = x(n)-\hat{s}(n)$

Empirically, the stability factor μ that controls the rate of convergence was determined to be 10^{-13}.

The starting set of B_0 coefficients was set to 0.077. Order of filter was equal 12. Algorithm was found to converge in the mean and remained stable under these conditions.

6.3.3 Fundamental Frequency Estimation

Pitch period (inverse of fundamental frequency F_0) was computed over frame lengths of 200 samples (25ms) using modified SIFT algorithm [1]. Fundamental frequencies in the desired range 50-250Hz were estimated. Speech samples were prefiltered by fifth order elliptic low-pass filter with f_{3dB} - 700Hz and 40dB attenuation for f=800Hz with transfer function as

$$H(z) = \frac{\sum_{i=0}^{N} P(i) z^{-i}}{\sum_{i=0}^{N} A(i) z^{-i}}$$

Filter coefficients are shown in Table 6.2.

i	A(i)	P(i)
0	1.00000000	0.02920214
1	-4.06679008	-0.05994938
2	7.03013315	0.03692716
3	-6.38120434	0.03692716
4	3.03280895	-0.05994938
5	-0.60260782	0.02920214

Table 6.2 Elliptic low-pass filter coefficients

Ripples of the obtained filter were no more than 1dB in bandpass and bandstop.

Down-sampling by 5 reduced effective sampling rate to 1.6kHz and reduced the overall number of further computations.

The autocorrelation of the low-pass filtered signal was calculated and the estimation of peaks of autocorrelation within desired pitch range was accomplished.

Parabolic approximation provided resolution $\geq \frac{5}{fs}$ (0.6ms). Estimated pitch periods varied from 5ms to 14ms. Approximately 10 percent of frames were classified as the unvoiced. The computer program used for pitch computation is submitted in App. C.

6.3.4 Results of Experiments With Cancelling of an Additive Wideband Noise and Telephone Noise

In this study, the experiments with noisy speech corresponding to signal-to-noise ratio values of 30, 20, 10 and 0dB were evaluated. The noise was generated according to the technique described in Chapter IV. Figures 5.2 and 6.3 up to 6.6 demonstrate the waveforms of a high quality speech, with additive noise 10 and 0dB, and with cancelled noise, respectively.

Looking at the waveforms of noise cancelled speech one easily can notice the smoothing effect of an algorithm. The improved quality of a voice is especially observed for

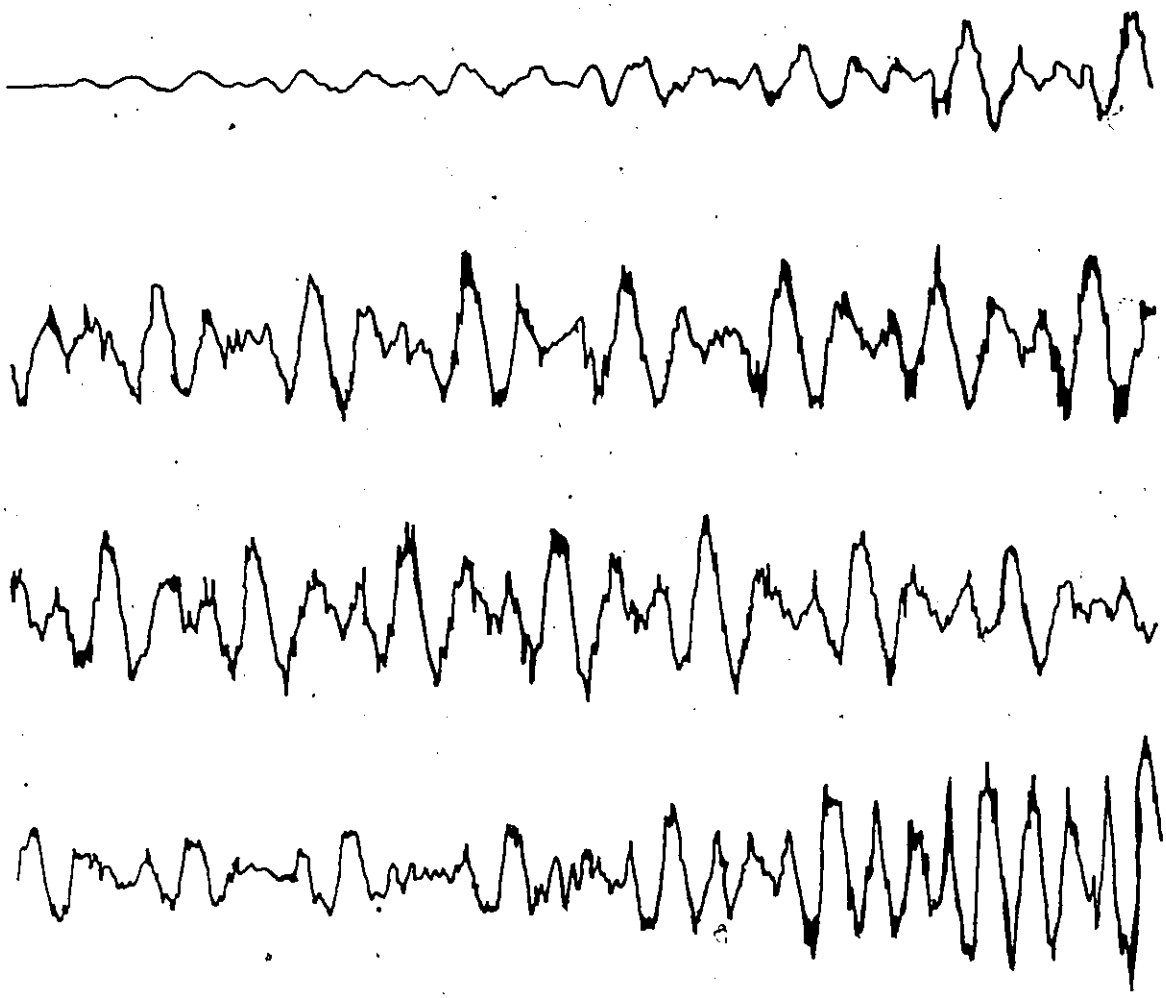


Fig. 6.3. The waveform of the speech with SNR=10dB.

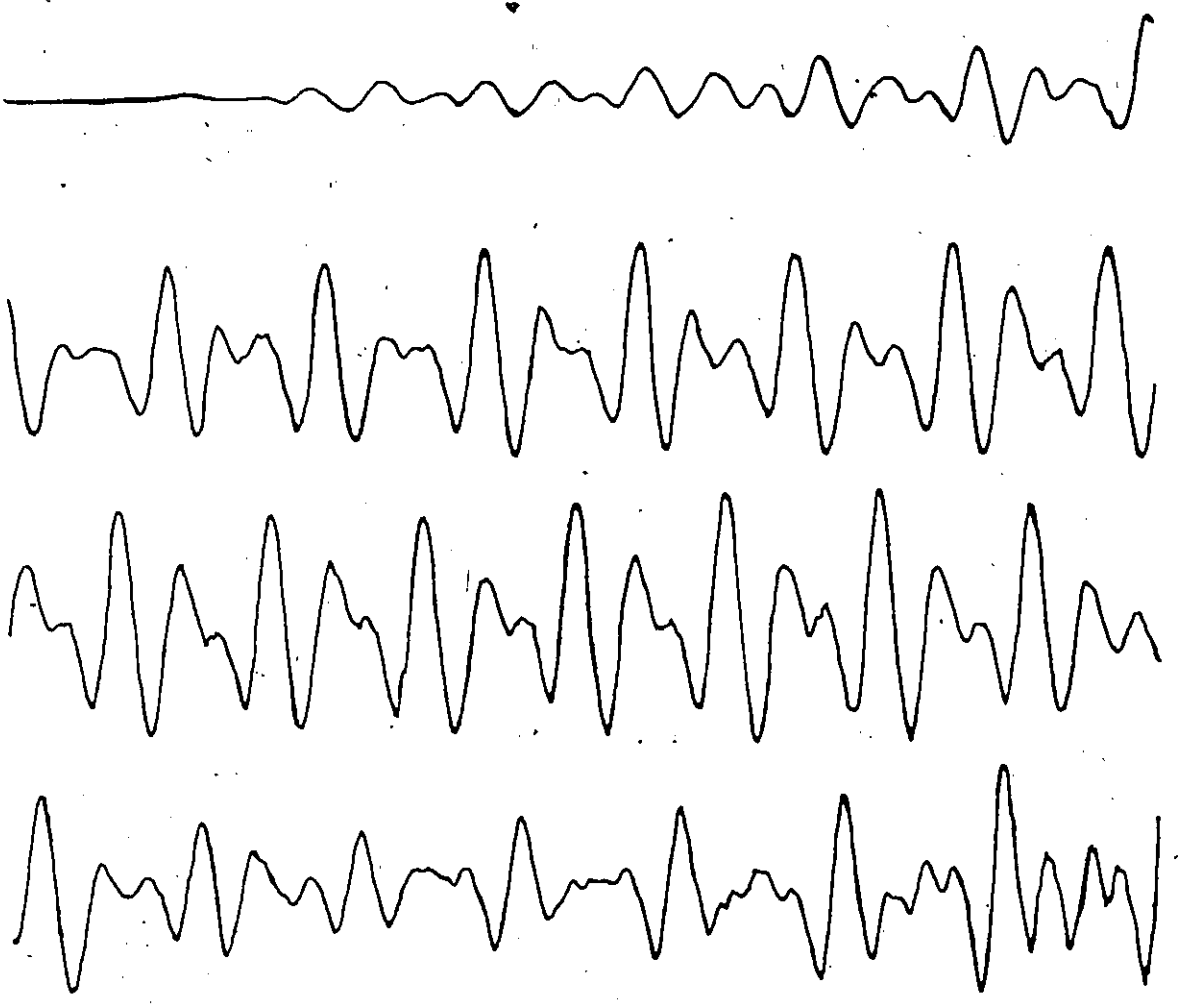


Fig. 6.4. The waveform of noise cancelled speech with SNR=10dB.

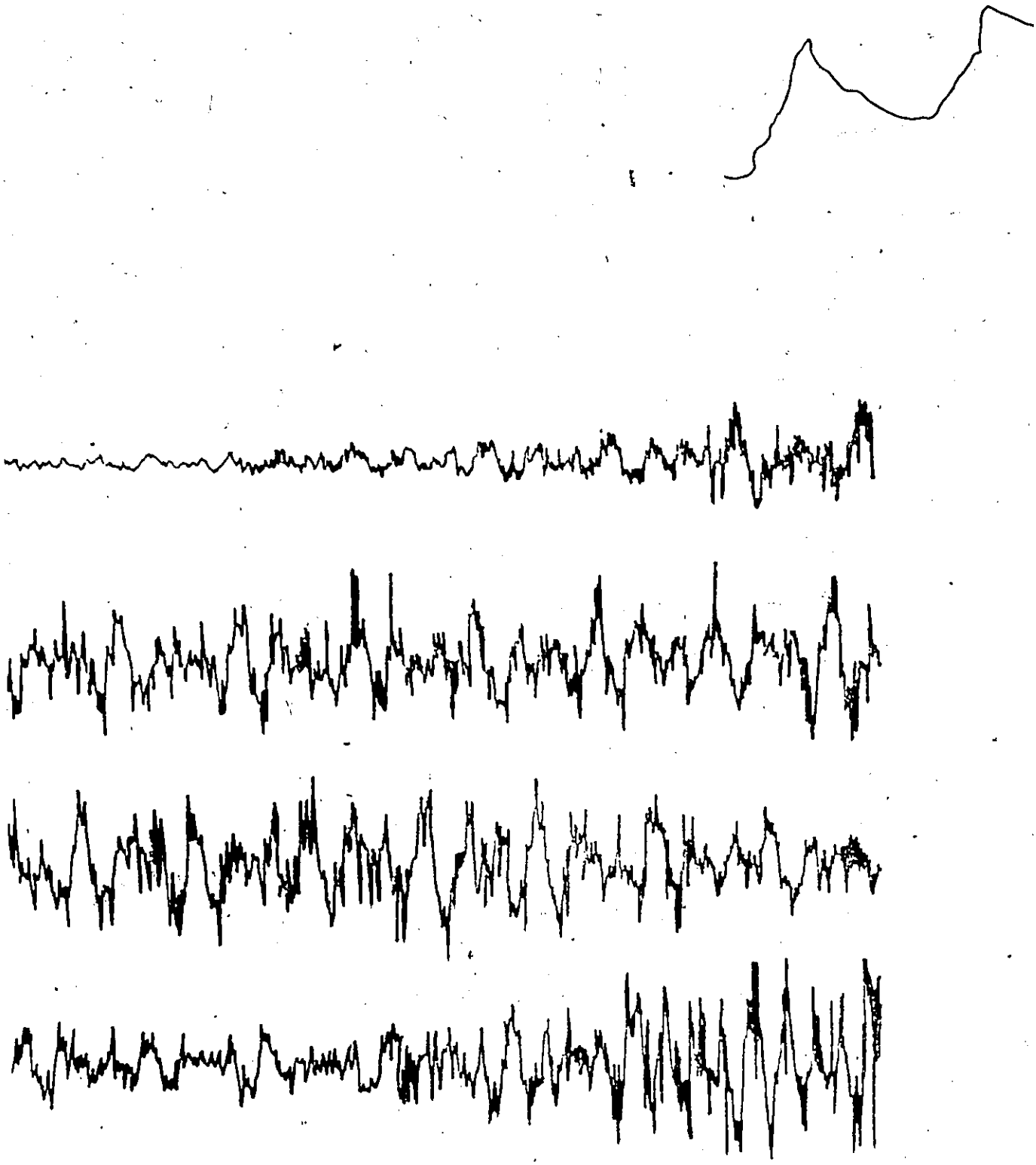


Fig. 6.5. The waveform of the speech with SNR=0dB.

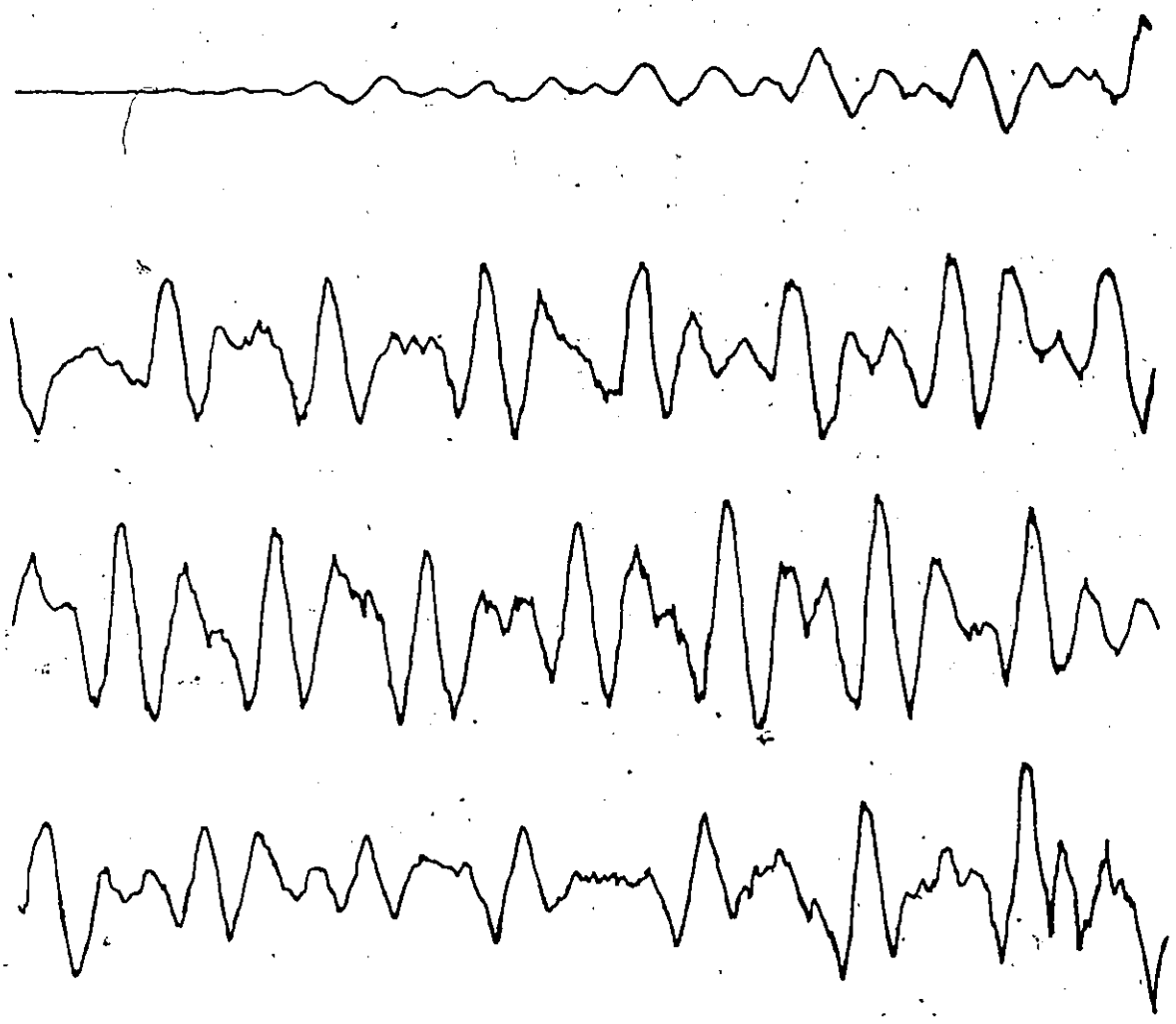


Fig. 6.6. The waveform of the noise cancelled speech with SNR=0dB.

the case where SNR was equal zero. The speech sounds less "harsh". It results in a significant improvement of a signal-to-noise ratio by 8.5dB.

For SNR more than 10dB, a perceptual improvement is unnoticeable.

We didn't make any extensive intelligibility tests of noise cancelled speech, because our main goal was just to improve the LPC analysis of a noisy speech.

Finally, the effectiveness of the adaptive filter was tested for noise removal from the telephone speech.

Waveforms of the original telephone speech and pre-processed one are shown in Figs 5.3 and 6.7. The noise cancelled telephone speech reminds closely the high quality speech as depicted in Fig. 5.2.

Distortions due to carbon microphone were almost eliminated. Perceptually both utterances sound very similarly and it is difficult to distinguish between them.

The effect of adaptive filtering on LPC analysis is illustrated in Fig. 6.8, where the periodograms of the LPC model of a phrase "We were away a year ago" spoken simultaneously through high quality dynamic microphone and through the telephone set are shown.

The dotted line represents the periodogram of the noise cancelled telephone speech. One can see the significant

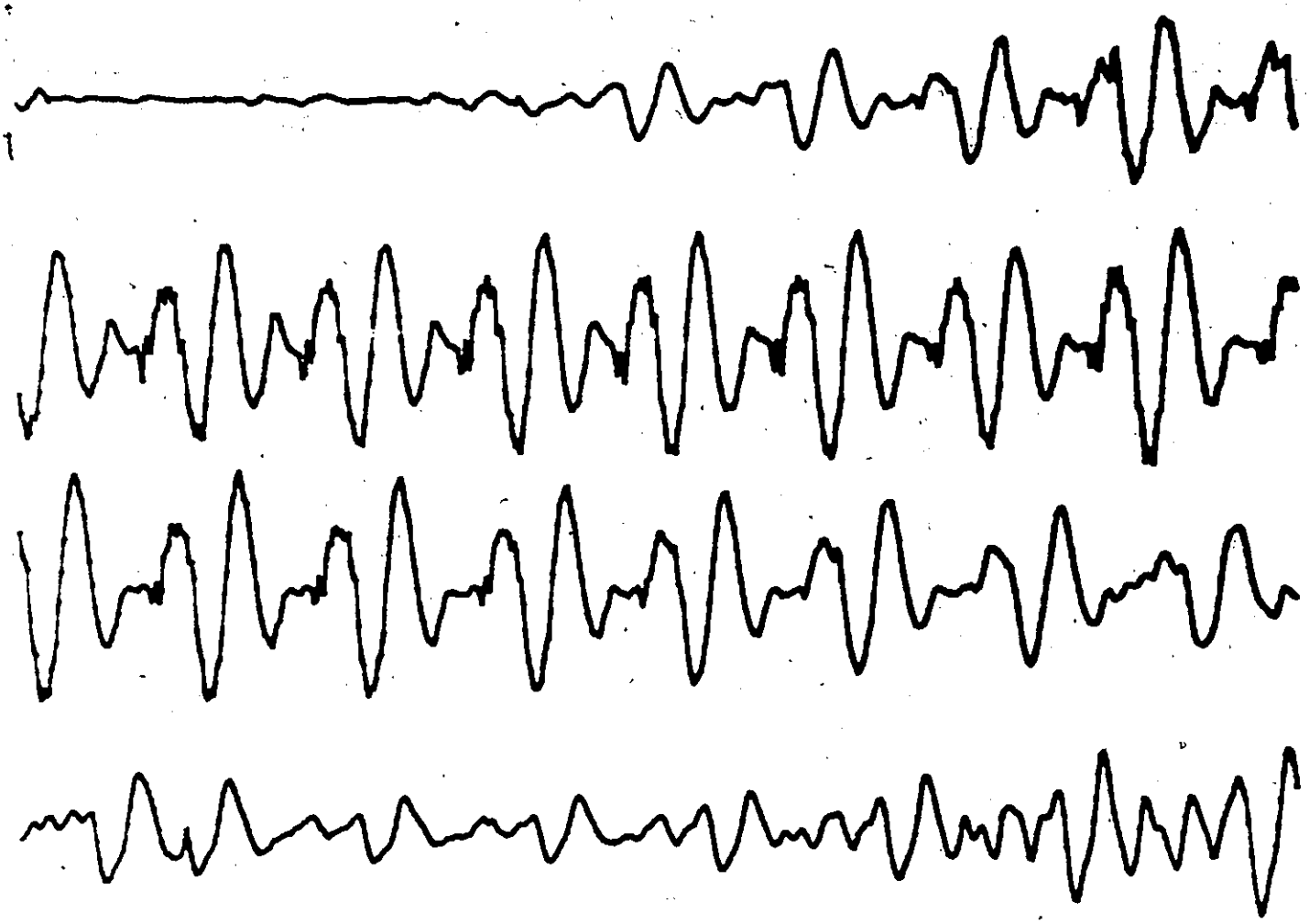


Fig. 5.7. The waveform of the noise cancelled telephone speech.

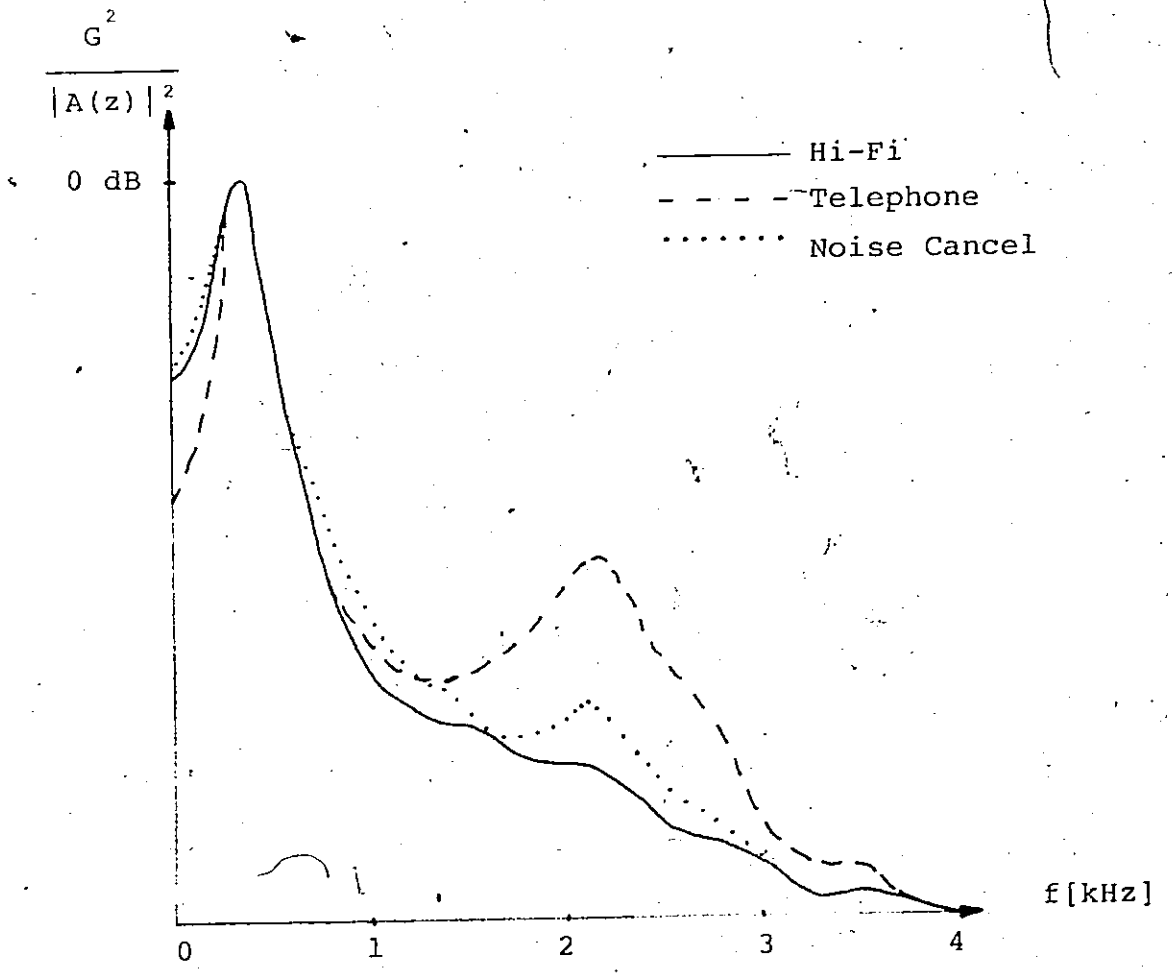


Fig. 6.8. Periodograms of the utterance, "We were away a year ago"

improvement in estimation of formants magnitude, and power distribution over the range 0-4kHz. The resulting periodogram of the LPC model is nearly similar to that one derived from high quality speech.

It should be added that typical dynamic range of a high quality signal, noise cancelled signal and telephone is equal to 32, 30 and 19.5dB, respectively.

Improvement in a signal-to-noise ratio was found to be 7-8dB over the original telephone speech.

6.4 VERIFICATION WITH PREPROCESSING

In a view of these results, a speaker verification procedure was applied to the preprocessed, noise cancelled telephone utterances. The talker dependent orthogonal parameters were calculated via reference eigenvectors derived previously from high quality speech.

They can be expressed as

$$P_{ij} = \sum_{k=1}^{N} V_{ki_{ref}} a_{kj}$$

where a_{kj} - a set of LPC parameters derived from noise cancelled speech

$v_{kj_{ref}}$ - reference eigenvectors obtained from high quality utterances

$N = 12$

The results of verification process are presented in Table 6.3.

The overall accuracy 96.6% was obtained when all the orthogonal parameters were used into the distance computation.

6.5 VERIFICATION TIME

The processing time for single verification is shown in Table 6.4. The duration of a test utterance is expressed here as a time unit T (1.5-2.5 sec.). The pitch detection analysis and noise cancelling algorithm employed in SVS operation increased significantly the verification time.

However, using subroutines written in assembly language and special purpose hardware the operation can be performed nearly in real time.

6.6 DISCUSSION

1) A technique for automatic speaker verification which deals with three kinds of speech

   a) high quality utterances

   b) the speech with additive pseudo-random wideband noise

   c) telephone utterances

was described.

2) Speaker verification accuracy from unprocessed telephone speech via references obtained from high quality speech was found to be as low as 45%.

Number of Orthogonal Parameters	Accuracy [%]
1	68.1
2	76.3
3	90.5
4	92.3
5	94.5
6	94.9
7	95.9
8	96.1
9	96.4
10	96.4
11	96.6
12	96.6

Table 6.3. Verification Accuracy of SVS from Noise-Cancelled Telephone Speech versus High Quality References.

Pitch Extraction	38T
Noise Cancelling	11T
LPC Analysis	22T
Eigenvector Analysis and Orthogonal Parameters	7T
Comparison and Decision	2T
Miscellaneous	10T
TOTAL	90T

Table 6.4. Processing Time for One Verification.

3) The adaptive noise cancelling algorithm has been found efficient to remove noise and distortions with unknown statistics from the telephone speech.

4) Orthogonal parameters from noise cancelled speech computed via reference eigenvectors derived from high quality speech have preserved the speaker verification capability.

5) The use of adaptive noise cancelling algorithm and 12^{th} order model of linear prediction makes it possible to realize high accuracy (>96%) verification when test utterances are subject to varying noise conditions.

6) Due to the need of only one set of references for a given speaker, the SVS memory requirements are reduced.

CHAPTER VII

SUMMARY AND CONCLUSIONS

Summary

In this work, the development of a Speaker Verification System that would operate under varying ambient conditions was investigated.

On the basis of a critical evaluation of the different techniques available for speaker verification the use of orthogonal parameters derived from the LPC model of speech was chosen as the primary verification technique. The evaluation of the effectiveness of the orthogonal parameter model was then undertaken for utterances obtained under different ambient conditons.

Methods for improving the accuracy of verification were then developed and tested. Specifically, techniques were developed to deal with the problem of satisfactory verification when:

i) speech utterances were relatively noise free

ii) speech utterances were subjected to additive wideband noise

iii) speech utterances were obtained over noisy telephone lines

iv) speech utterances were obtained under unspecified but varying ambient conditions.

Conclusions

The conclusions based on the work described above can be briefly stated as follows:

i) SVS for high quality speech:

The 8th order of linear prediction was found to be sufficient to ensure high verification accuracy. It was concluded that two least significant orthogonal parameters are the most speaker dependent features and only they were used in measurements of dissimilarity between speakers.

Experiments have shown the time normalization procedure was essential for high accuracy of verification algorithm. It may mean, that even the least significant orthogonal parameters are not completely free of linquistic content of the speech. The concept of the minimum rejection-to-acceptance ratio was found useful to set an appropriate threshold.

ii) SVS for noisy speech:

The results of the original study on sensitivity of orthogonal parameters to noise have shown, that the last few orthogonal parameters are the least sensitive to additive noise and still preserve the speaker verification capability. Experiments indicated that the 12th order of linear prediction and last three orthogonal parameters make the SVS able to maintain the high verification accuracy even with SNR as low as 12 dB.

The expression was derived in order to determine the orthogonal parameters of a clean signal in terms of para-

meters derived from noisy signal. Moreover, the mutual relation between the LPC parameters of clean and noisy speech was derived.

iii) SVS over telephone lines:

The feasibility of such a system has been demonstrated. The sources of noise and distortions were analyzed and spectral analysis of signal and noise in a telephone line was performed. The error rate was around 1%.

iv) Flexible SVS:

A new concept of a SVS which would operate with different speech inputs and rely on only one set of references, was introduced and developed.

Results of preliminary studies with unprocessed noisy speech have indicated that the speaker verification based on high quality speech references has very low accuracy.

An adaptive noise cancelling technique applied to telephone speech gave promising results. The similarity of the spectral estimations for both high quality and noise cancelled speech has underlined the significant improvement on linear prediction analysis performed of noisy speech.

The overall error rate of the verification procedure was found to be less than 4% when the SVS operated on noise cancelled telephone speech.

The verification accuracy can be potentially improved by eliminating the unvoiced frames, from further processing and concentrating only on voiced parts of speech.*

The results of this dissertation have shown, that orthogonal parameters may be used as reliable speaker-specific features in SVS algorithms under relatively different noise conditions.

There are some suggestions for future work:

a) hardware implementation of the automatic speaker verification system operating in real time using an array processor would be desired;

b) further investigations are needed to test the performance of the SVS based on orthogonal parameters for attempts of mimics, twins and uncooperative speakers;

c) applicability and efficiency of this algorithm for reliable text-independent speaker verification.

---

* As we remember, in noise cancelling algorithm during frames classified as unvoiced the output simply followed the noisy input. This potentially introduces the error in further LPC analysis.

APPENDIX A

PARAMETRIC REPRESENTATION OF SPEECH PRODUCTION MODEL

Speech signals are composed of a sequence of sounds. They are produced as a result of acoustical excitation of the human vocal tract. An acoustic pressure wave is expelled from the lungs into trachea and then forced between the vocal folds. The opening between the vocal folds is called the glottis.

The rate of the vocal folds vibrations determines the fundamental frequency or pitch of the voice. During the production of voiced sounds, the vocal tract is excited by a series of quasi-periodic pulses of air generated by the vocal cords adjusted so that they vibrate in a relaxation oscillation. Unvoiced sounds are generated, when the excitation is provided by air passing turbulently through constrictions in the vocal tract. The vocal tract can be represented as a non-uniform acoustic tube which consists of a set of interconnected sections of equal length and extends from the glottis to the lips. Each individual section is of uniform area.

The transverse dimension of each section is small enough compared with a wavelength so that the sound propagation through an individual section can be treated as a plane wave.

Different sounds are formed by varying the shape of the vocal tract. Approximate model of the vocal tract can be made by representing it as a discrete time-varying linear filter. If the variations with time of the vocal tract shape are assumed to be approximated with sufficient accuracy by a succession of stationary shapes, it is possible to define a transfer function of the filter in the complex Z-domain.

The linear speech production model described in Z-transform notation can be expressed by the following equation:

$$S(z) = E(z)G(z)V(z)L(z) \qquad (A.1)$$

where $S(z) \leftrightarrow s(n)$ - speech signal

$E(z) \leftrightarrow e(n)$ - excitation function

$G(z) \leftrightarrow g(n)$ - glottal shaping

$V(z) \leftrightarrow v(n)$ - vocal tract

$L(z) \leftrightarrow l(n)$ - lip radiation

A sequence of unit pulses, spaced over the pitch period represents the excitation function E(z)

$$E(z) = D \sum_{n=0}^{\infty} (z^{-K})^n = \frac{D}{1-z^{-K}} \qquad (A.2)$$

for $|z|>1$

where K - number of sampling periods within the pitch period.

The glottis can be modeled by the equation

$$G(z) = 1/(1-e^{-cT}z^{-1})^2 \qquad (A.3)$$

which represents the two-pole low-pass filter with an estimated cutoff frequency around 100Hz. Both poles are on the real axis inside the unit circle.

The lip radiation effect is described by the first order high-pass filter.

$$L(z) = 1-z^{-1} \qquad (A.4)$$

The vocal tract can be modeled by the cascade of K second order all-pole filters

$$V(z) = \frac{1}{\prod_{i=1}^{K} [1-2e^{-c_i T}\cos(b_i T)z^{-1}+e^{-2c_i T}z^{-2}]} \qquad (A.5)$$

where K - a number of formants defined in the model

$F_i = \frac{b_i}{2\pi}$ -- the i^{th} formant frequency

$B_i = \frac{c_i}{2\pi}$ -- the bandwidth of the i^{th} formant.

The zero which reflects the lip radiation effect can be easily cancelled by one of the poles of glottal shaping, because the term cT is generally much less than unity.

In addition, the location of a pole is considerably more important perceptually than the location of a zero.

In most cases the zeros (antiresonances) contribute only to the spectral balance. Thus, an explicit representation of the antiresonances by zeros of the linear filter is not necessary. An all-pole model of the vocal tract can approximate the effect of antiresonances on the speech wave in the frequency range of interest to any desired accuracy.

The total form of speech production model can be then expressed by the following formula:

$$S(z) = \frac{E(z)}{(1-e^{-cT}z^{-1}) \prod_{i=1}^{K} [\, 1-2e^{-c_i T}\cos(b_i T)z^{-1}+e^{-2c_i T}z^{-2}]} \qquad (A.6)$$

The transfer function of this recursive filter is defined as

$$A(z) \simeq \frac{1}{G(z)V(z)L(z)} = \sum_{i=0}^{M} a_i z^{-i} \qquad (A.7)$$

where $a_0 = 1$

$M \geq 2K+1$ -- order of filter.

The synthesis and analysis models are expressed as:

$$S(z) = E(z)\frac{1}{A(z)} \qquad (A.8)$$

and

$$E(z) = S(z)A(z) \qquad (A.9)$$

One of the important factors of these models is that the combined contributions of the glottis, the vocal tract, and the lip radiation are represented by a single recursive filter.

The order M of this filter required to represent the speech adequately is determined by the number of resonances and antiresonances of the vocal tract in the frequency range of interest, the nature of glottal shape and the radiation.

The number of filter coefficients M in the filter equals the number of sections M in the acoustic tube model. The relation between sampling frequency f_s, number of sections M, length of the acoustic tube L, and speed of sound c is given by the equation

$$f_s = \frac{Mc}{2L} \qquad \text{(A.10)}$$

where $T = 2l/c$ - sampling period

L - length of the vocal tract (approximately 17 cm.)

The linear predictor memory must be equal to twice the time required for sound waves to pass from the glottis to the lips. It takes about 1ms for the sound to travel 34cm, hence for example for f_s=8kHz, the number M should be at least 8, to reflect the properties of the vocal

tract. Additional one or two poles for glottis make usually the number M larger. It should be added, that this is only a rough estimate of M and it depends to certain degree on the speaker and the linguistic content of the speech.

APPENDIX B

Linear Prediction Theory in Speech

The first term "linear prediction" appeared in Wiener's book "The linear predictor for a single time series" published in 1949. In speech domain, it refers to a representation of the voice signal as the output of an all-pole time-varying digital filter that is excited by a sequence of pulses spaced by the pitch period for voiced sounds, or random noise for unvoiced sounds.

The prediction theory says, that the actual sample of voice signal can be predicted from M previous samples multiplied by appropriate weights coefficients. The output of the linear filter at the nth sampling instant of time is given by

$$s(n) = \sum_{k=1}^{M} a_k s(n-k) + Ku(n) \qquad \text{(B.1)}$$

where a_k - are the linear prediction coefficients (LPC) that account for the filtering action of the vocal tract, the radiation, and the glottis.

$u(n)$ represents the nth sample of the appropriate excitation

M - is the order of prediction

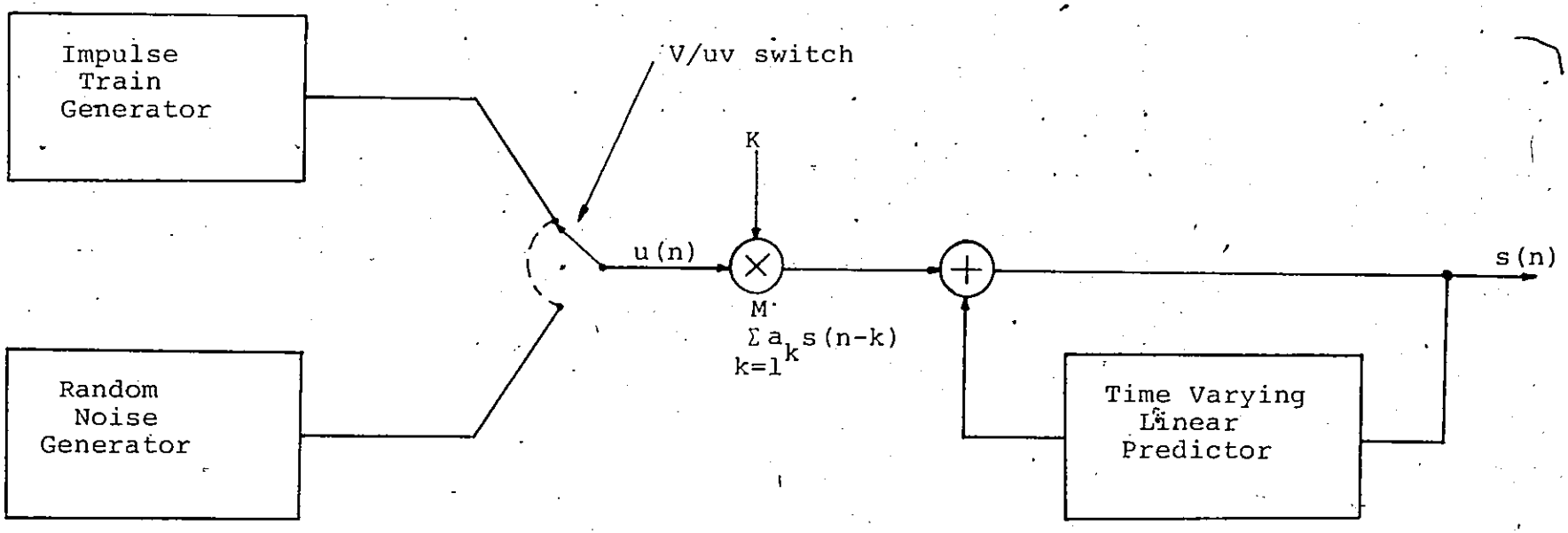
The transfer function of the linear, all-pole filter is expressed by

$$H(z) = \frac{K}{1-\sum_{k=1}^{M} a_k z^{-k}} \qquad (B.2)$$

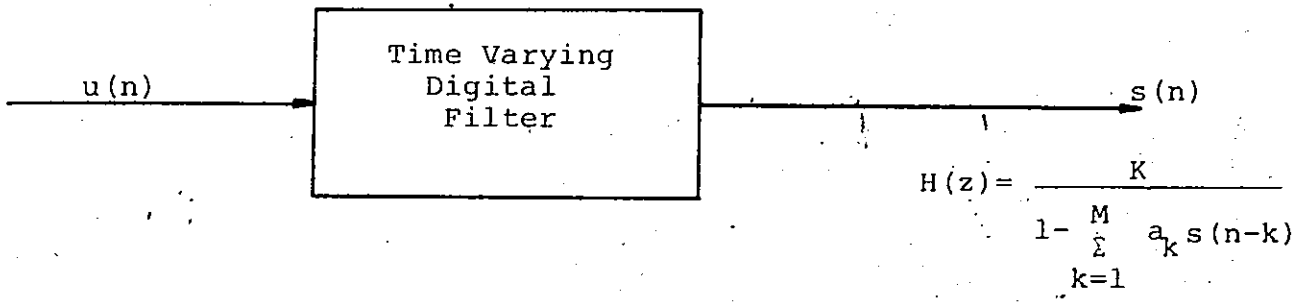
The time domain as well as the equivalent frequency domain representation of linear prediction model of speech are shown in Fig. B.1.

The parameters of this model like the prediction coefficients a_k, the pitch period T, the rms value of the speech samples, and a binary parameter indicating whether the speech is voiced or unvoiced, provide a complete representation of the speech waveform over a short time period during which the vocal tract shape is assumed to be constant. In practice, of course, the vocal tract changes continuously in time. However, in most cases, to account for the nonstationary character of the speech wave, the predictor coefficients a_k of the modeled filter are periodically readjusted, once every 10 to 30 ms.

The simplified all-pole model is a natural representation of non-nasal voiced sounds, but for nasals and fricative sounds, the detailed acoustic theory calls for both poles and zeros in the vocal tract transfer function.



a) Time domain



b) Frequency domain

Fig. B.1. All-pole model of linear prediction speech production.

However, if the order of filter is chosen to be high enough, the all-pole model provides a good representation for almost all the sounds of speech.

The basic problem of linear prediction analysis is to determine a set of predictor coefficients a_k directly from the speech signal.

First, the prediction error function e(n) is defined as a difference between the true value of the nth sample of the signal and the predicted one.

(B.3) $$e(n) = s(n) - \hat{s}(n)$$

where $\hat{s}(n)$ is the predicted sample of the speech expressed by

$$\hat{s}(n) = \sum_{k=1}^{M} a_k s(n-k)$$

The basic problem of linear prediction analysis is to determine a set of linear prediction coefficients a_k which minimize the error function e(n). The importance of linear prediction is that the LPC parameters characterizing the time-varying filter can be determined directly from the speech waveform by applying a least squares criterion. The optimal set of filter coefficients is obtained by minimization of the sum of the squares of some specified number of error samples.

The total squared prediction error E over the frame of speech is defined by

$$E = \sum_{n} e^2(n) \tag{B.5}$$

Since the error function e(n) can be expressed by

$$e(n) = \sum_{k=0}^{M} a_k s(n-k) \tag{B.6}$$

hence

$$E = \sum_{n} \left[\sum_{k=0}^{M} a_k s(n-k) \right]^2 = \tag{B.7}$$

$$= \sum_{n} \sum_{k=0}^{M} \sum_{l=0}^{M} a_k s(n-k) s(n-l) a_l$$

The values of a_k are found by setting the partial derivation of E to zero, for i=1,2,3....M, and solving a set of M simultaneous equations.

$$\frac{\delta E}{\delta a_i} = 2 \sum_{n} \sum_{k=0}^{M} a_k s(n-k) s(n-i) = 0 \tag{B.8}$$

Defining

$$\psi_{ki} = \sum_{n} s(n-k) s(n-i) \tag{B.9}$$

and since the coefficient a_o equals to one, hence

$$\sum_{k=1}^{M} a_k \psi_{ki} = -\psi_{oi} \tag{B.10}$$

Solving this set of M linear simultaneous equations, the M coefficients a_k are computed.

In general, the solution of a set of simultaneous equations requires a great deal of computation. However, since the matrix of coefficients ψ_{ki} is symmetric and positive defined, the number of required operations can be substantially reduced. Sometimes, the linear prediction coefficients a_k obtained by solving Eq. B.10 generate poles of the filter transfer function outside the unit circle. This can happen whenever a pole located very close to the unit circle appears outside the unit circle due to approximations in the model or errors caused by truncation or round-off operation in computations.

There are several methods of solving a set of equations B.10. They differ basically in the approach to the limits of summation over the total squared prediction error E(n) is calculated, and the definition of analyzed speech sequence s(n).

The covariance method and the autocorrelation method are the most frequently used techniques in linear prediction analysis.

Assuming, that the speech sequence is defined as $\{s(0),s(1)....s(N-1)\}$, the covariance method requires that the prediction error function is minimized over finite

interval of time (M,N-1), and all N speech samples are included into calculations of the covariance matrix ψ_{ki} i.e.,

$$e(n) = \sum_{k=0}^{M} a_k s(n-k) \qquad \text{(B.11)}$$

where N=M, M+1....N-1

and

$$E = \sum_{n=M}^{N-1} \left[\sum_{k=o}^{M} a_k s(n-k) \right]^2 \qquad \text{(B.12)}$$

It results in a set of M-simultaneous equations

$$\sum_{k=1}^{M} a_k \psi_{ki} = -\psi_{oi} \qquad \text{(B.13)}$$

$$\text{where } \psi_{ki} = \sum_{n=M}^{N-1} s(n-k)s(n-i) \qquad \text{(B.14)}$$

and i=1,2,3....M

In matrix form the equations B.13 become

$$\begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} & \cdots & \psi_{1M} \\ \psi_{21} & \psi_{22} & \psi_{23} & \cdots & \psi_{2M} \\ \psi_{31} & \psi_{32} & \psi_{33} & \cdots & \psi_{3M} \\ \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \\ \cdots & \cdots & \cdots & \cdots & \\ \psi_{M1} & \psi_{M2} & \psi_{M3} & \cdots & \psi_{MM} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ \cdots \\ \cdots \\ a_M \end{bmatrix} = - \begin{bmatrix} \psi_{01} \\ \psi_{02} \\ \psi_{03} \\ \cdots \\ \cdots \\ \cdots \\ \psi_{0M} \end{bmatrix}$$

This matrix is symmetric, but the diagonal elements are not equal. They are related by the equation

$$\psi_{k+1,i+1} = \psi_{k,i} + s_{-i-1} \cdot s_{-k-1} - s_{N-1-i} \cdot s_{N-1-k} \tag{B.16}$$

The name "covariance method" is derived from the fact, that for zero mean signal, the elements of matrix Φ define the covariance matrix.

The autocorrelation method is the second approach to determining the linear prediction coefficients. In this method, the total squared error given by Eq. B.7 is minimized over the infinite interval $-\infty<n<\infty$.

However, the speech signal $s(n)$ is windowed to be zero outside the interval $0 \leq n \leq N-1$, using the rectangular or Hamming window of finite length. The equation B.9 can be simplified as follows

$$\psi_{ki} = \sum_{n=-\infty}^{\infty} s(n-k)s(n-i) = \sum_{n=0}^{N-1-|k-i|} s(n)s(n+|k-i|) = r(|k-i|) \tag{B.17}$$

As a result ψ_{ki} reduces to the short-term autocorrelation $r(t)$ at the delay $\ell=|k-i|$.

The set of M simultaneous equations is given by

$$\sum_{k=1}^{M} a_k r(|k-i|) = -r(i) \tag{B.18}$$

where

$$r(|k-i|) = \sum_{n=0}^{N-1-|k-i|} s(n)s(n+\ell)$$

$$i = 1,2,3 \ldots\ldots M$$

The set of equations (B.18) can be expressed in matrix form as

$$
\begin{bmatrix}
r(0) & r(1) & r(2) & \cdots & r(M-1) \\
r(1) & r(0) & r(1) & \cdots & r(M-2) \\
r(2) & r(1) & r(0) & \cdots & r(M-3) \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
r(M-1) & r(M-2) & r(m-3) & \cdots & r(0)
\end{bmatrix}
\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ a_M \end{bmatrix}
=
\begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \cdot \\ \cdot \\ r(M) \end{bmatrix}
\tag{B.20}
$$

The MxM matrix of autocorrelation is a Toeplitz matrix; i.e., it is symmetric and all the elements along a given diagonal are equal.

Hence, only M(M+1)/2 elements need to be determined, there are several efficient recursive procedures for solving this system of equations, as Choleski decomposition[1], Levinson's method[1], Robinson's method[1], Durbin's recursive solution[40].

The relationship between a nonuniform acoustic tube formed by cascading M uniform cylindrical sections and the linear prediction model enables the direct representation of LPC parameters a_k in terms of reflection parameters (PARCOR) μ_k, and the area coefficients A_k.

The area coefficients are uniquely related to the reflection coefficients by

$$\frac{A_k}{A_{k-1}} = \frac{1-\mu_k}{1+\mu_k} \quad \text{, for } 1 \leq k \leq M \quad \text{where } |\mu_k| < 1 \text{ and } A_{M+1} = 1 \tag{B.21}$$

The relation between parameters a_k and reflection coefficients μ_k is given by

$$\mu_k = a_k^{(k)} \quad \text{, } k=1, 2....M \tag{B.22}$$

where $a_k^{(i)}$ represents the k^{th} LPC for an i^{th} pole linear prediction model.

Since the reflection coefficients are calculated in autocorrelation method as a by-product, and are bounded by unity from the definition, they can be used as a natural built-in stability test for the filter $1/A(z)$. Theoretically, the autocorrelation method assumes the stability for $|\mu_k| \leq 1$, $k=1, 2....M$.

This is an important advantage over the covariance method, where certain problems with the stability can arise. For example, if the inverse filter, $A(z)$, has a root on or outside the unit circle, the unit sample response of $1/A(z)$ will have infinite energy. If this filter is unstable, the results obtained by further processing, e.g., generation of an acoustic tube for parameter

quantization, performing the spectral analysis, etc., must be carefully interpreted.

There is a way to replacing all roots outside the unit circle by their reciprocals, which will assure stability [2], but the original linear prediction formulation is no longer being solved.

As a general statement, the autocorrelation method for linear prediction analysis is global in that several pitch periods must be contained within the analysis window for meaningful results with voiced sounds, i.e. $N>>M$.

On the other hand, the covariance method can be used to intervals less than pitch period 5-10ms. For unvoiced speech, both methods give similar results for frames longer than 10ms. The similarity of the final results is caused by the fact that when the number of samples N is sufficiently large, the covariance coefficients ψ_{ki} are approximately equal to the autocorrelation coefficients r(k-i).

The approximate number of operations necessary to implement the autocorrelation method is less than for the covariance method [1].

For typical case N=128, M=10, the total number of operations (additions and multiplications) for matrix load

and matrix solve using the covariance and autocorrelation method was found to be 2080 and 1427 respectively. It results in 1.4 times longer computations in covariance method. Moreover, the direct application of standard linear simultaneous equation solution and direct coefficient evaluation, would require approximately six times greater computational effort for the covariance method.

In conclusion, the autocorrelation method is faster, is assured of being stable, and allows a meaningful spectral matching gain to be computed. As a result, it would seem to be the preferable choice and consequently, it was chosen in our experimentations.

APPENDIX C

Computational Details

```
C          "ANCOEF"
C
C       PROGRAM COMPUTES LINEAR PREDICTION ,REFLECTION & AUTOCORELATION
C               COEFFICIENTS
        DIMENSION NAME(5),IX(300),A(16),R(16),X(300)
        ACCEPT "# OF LPCOEF =",NPP
        ACCEPT '# OF FRAMES =',NS
        NP=NPP
        NP1=NP+1
        ICOUNT=0
    5   ACCEPT '# FOR ANALYSIS =',NFT
        ND=NFT
        TYPE "ENTER OUTPUT FILENAME"
        READ(11,1)(NAME(I),I=1,5)
    1   FORMAT(5A2)
        OPEN 1,NAME
        TYPE "ENTER INPUT FILENAME"
        READ(11,1)(NAME(I),I=1,5)
        OPEN 0,NAME
        NFT=ND/NS
        DO 11 K=1,NS
        ICOUNT=ICOUNT+1
        IF(K.EQ.NS)NFT=ND-(K-1)*NFT
        READ(0)(IX(J),J=1,NFT)
        DO 2 J=1,NFT
    2   X(J)=IX(J)/50.0
        CALL AKN1(X,A,R,NP,NFT)
        Z=NS
        WRITE(1) Z,  (A(J),J=1,NP)
C       WRITE(12,10) (A(J),J=1,NP)
 10     FORMAT(2(" ",2(6E14.7,/ )))
C        WRITE(12,15) ICOUNT
  15     FORMAT("0"2X,"# OF FRAME",I5)
   11   CONTINUE
        CLOSE 1
        CLOSE 0
        ACCEPT 'WISH TO CONTINUE ? YES-1,NO-0 ',IWISH
        IF(IWISH.EQ.1)GO TO 5
        END
```

```
      SUBROUTINE AKN1(X,A,R,NP,L,AL)
C       L-NUMBER OF SPEECH SAMPLES IN FRAME
C       X-SPEECH DATA
        A-LPC PARAMETERS
        R-AUTOCORRELATION COEFFICIENTS
       CK-REFLECTION COEFFICIENT1
C       AL-GAIN
C       NP-LINEAR PREDICTION ORDER
      DIMENSION X(L)
      DIMENSION A(16),CK(20),R(16)
      IP=NP+1
      DO 1 J=1,IP
    SUM=0.
    KK=L-J+1
    DO2K=1,KK
    SUM=SUM+X(K)*X(K+J-1)
  2 CONTINUE
    R(J)=SUM
  1 CONTINUE
      A(1)=-R(2)/R(1)
      CK(1)=A(1)
      AL=R(1)+A(1)*R(2)
      BE=R(3)+A(1)*R(2)
      MM1=NP-1
      DO10I=1,MM1
      IP1=I+1
      CC=-BE/AL
      CK(IP1)=CC
      I2=IP1/2
      DO20J=1,I2
      IJ=IP1-J
      TA=A(J)+CC*A(IJ)
      A(IJ)=A(IJ)+CC*A(J)
 20   A(J)=TA
      A(IP1)=CC
      AL=AL+CC*BE
      BE=R(I+3)
      DO10J=1,IP1
      NJ=I-J+3
 10   BE=BE+A(J)*R(NJ)
      RETURN
      END
```

```
C     ********** GAUSS **********
      SUBROUTINE GAUSS(IX,S,AM,Y)
      A=0.0
      DO 50 I=1,12
    1 CALL RANDU(IX,IY,Y)
      IX=IY
 50   A=A+Y
      Y=(A-6.0)*S+AM
      RETURN
      END
```

```
      SUBROUTINE AKN2(X,A,R,NP,L,AL,VAR)
      DIMENSION X(L)
      DIMENSION A(16),CK(20),R(16)
      IP=NP+1
      DO 1 J=1,IP
      SUM=0.
      KK=L-J+1
      DO2K=1,KK
      SUM=SUM+X(K)*X(K+J-1)
    2 CONTINUE
      R(J)=SUM
    1 CONTINUE
      DO 5 I=1,IP
      WRITE(12,6) R(I)
    5 CONTINUE
    6 FORMAT(6E18.8)
C      INTRODUCTION OF VARIANCE OF NOISE TO R(1)
      TYPE "R(1)=",R(1),   "R(2)=",R(2)
      R(1)=R(1)-L*VAR
      WRITE(12,7) R(1)
    7  FORMAT(3X,E18.8)
      TYPE "NEW R(1)=",R(1)
      A(1)=-R(2)/R(1)
      CK(1)=A(1)
      AL=R(1)+A(1)*R(2)
      BE=R(3)+A(1)*R(2)
      MM1=NP-1
      DO10I=1,MM1
      IP1=I+1
      CC=-BE/AL
      CK(IP1)=CC
      I2=IP1/2
      DO20J=1,I2
      IJ=IP1-J
      TA=A(J)+CC*A(IJ)
      A(IJ)=A(IJ)+CC*A(J)
   20 A(J)=TA
      A(IP1)=CC
      AL=AL+CC*BE
      BE=R(I+3)
      DO10J=1,IP1
      NJ=I-J+3
   10 BE=BE+A(J)*R(NJ)
      RETURN
      END
```

```
C          "ORTPAR"
C
C      PROGRAM COMPUTES - THE REFERENCE COVARIANCE MATRIX OF LPC PARAMETERS
C                       - EIGENVECTORS AND EIGENVALUES
C                       - ORTHOGONAL PARAMETERS
C      INPUT DATA: LPC PARAMETERS
C      SUBROUTINE REQUIRED: EIGEN
       DIMENSION NAME(5),B(16,125),RR(16,16),BB(16),A(256),R(256),RX(16,16),
     *  X(16),EIG(16),E(16,16),ORT(16,125),AVERQ(16),AVORT(16)
       REAL SQRT,ABS,FLOAT
   50  ACCEPT"THE ORDER OF LPC MODEL=",NP
       TYPE "WISH TO FIND AND STORE EIGENVALUES ONLY ? III=1"
       TYPE "WISH TO  FIND AND STORE REF.EIGENVECTORS ? III=2 "
       TYPE "WISH TO FIND ORTHOGONAL PAR. VIA REF. EIGENVECTORS OF A
     *  GIVEN SPEAKER ? III=3 "
       TYPE "WISH TO FIND REFERENCE ORT. PAR. FOR A GIVEN SPEAKER? III=4"
       ACCEPT "III=?",III
       ACCEPT "LENGHT OF INPUT FILE=",NS
       NS=NS/4
       NF=NS/(NP+1)
       LAM=0
  220  MSUM=0
       ILL=0
       NN=0
       NA=NP*(NP+1)/2
       NR=NP*NP
       DO 15 I=1,NP
       DO 15 J=1,NP
       RX(I,J)=0.0
   15  AVERQ(I)=0.0
   55  IF(ILL.EQ.1) GO TO 34
       MSUM=MSUM+NF
       NN=NN+1
   34  TYPE "ENTER INPUT FILENAME"
       READ(11,5) (NAME(I),I=1,5)
    5  FORMAT(5A2)
       OPEN 0,NAME,LEN=4*(NP+1)
       DO 20 J=1,NF
       READ(0) (X(K),K=1,NP+1)
       DO 20 K=1,NP+1
       IF(K.EQ.1) GO TO 20
       B(K-1,J)=X(K)
   20  CONTINUE
       CLOSE 0
       IF(III.EQ.3) GO TO 999
```

```
C
C       CREATION OF REFERENCE COVARIANCE MATRIX FOR A GIVEN SPEAKER
C
        AVEB=0.0
        DO 1 I=1,NP
        DO 2 J=1,NF
        AVEB=AVEB+B(I,J)
    2   CONTINUE
        BB(I)=AVEB/NF
        AVEB=0.0
    1   CONTINUE
        DO 3 I=1,NP
        DO 3 K=1,NP
        SUMA=0.0
        DO 4 J=1,NF
        SUMA =SUMA+(B(I,J)-BB(I))*(B(K,J)-BB(K))
    4   CONTINUE
        RR(I,K)=SUMA/(NF-1)
    3   CONTINUE
        DO 33 I=1,NP
        DO 33 K=1,NP
   33   RX(I,K)=RX(I,K)+NF*RR(I,K)
        IF(III.NE.1) GO TO 155
        ACCEPT "WISH TO READ ANOTHER FILE, YES-1,NO-0",JA
        IF(JA-1)155,55,155
C
  155   L=1
        DO 45 K=1,NP
        DO 45 I=1,NP
        RX(I,K)=RX(I,K)/MSUM
        IF(I.GT.K) GO TO 45
        A(L)=RX(I,K)
        TYPE A(L)
        L=L+1
   45   CONTINUE
C
C       CALCULATION OF EIGENVALUES A(L) AND EIGENVECTORS R(L)
C
        CALL EIGEN(A,R,NP,NA,NR,0)
        DO 211 I=1,NP
        DO 211 K=1,NP
        L=(I-1)*NP+K
        E(K,I)=R(L)
        WRITE(12,6) E(K,I)
  211   CONTINUE
        IF(III.EQ.1) GO TO 889
        IF(III.EQ.2) GO TO 889
        IF(III.EQ.3) GO TO 999
        IF(III.EQ.4) GO TO 105
  889   ACCEPT "WISH TO STORE REF. EIGENVECTORS ? YES-1,NO-0",LALA
        IF(LALA.NE.1) GO TO 777
        TYPE "ENTER INPUT FILENAME "
        READ(11,5) (NAME(I),I=1,5)
        NPN=NP*NP
```

```
      OPEN 3,NAME,LEN=4*NPN
      WRITE (3) (R(L),L=1,NPN)
      CLOSE 3
777   WRITE(12,77)
 77   FORMAT("0","NEXT SET OF ORT-S")
      L=1
      DO 10 K=1,NP
      EIG(K)=A(L)
      TYPE EIG(K)
      L=L+K+1
 10   CONTINUE
  6   FORMAT(6E15.7)
      MEAN=MSUM/NN
      ACCEPT" WISH TO STORE REFERENCE FILE WITH EIGENVALUES AND AVERAGE # OF
     * FRAMES-? YES-1,NO-0",IND
      IF(IND.NE.1) GO TO 105
      TYPE "ENTER OUTPUT REFERENCE FILENAME"
      READ(11,5) (NAME(I),I=1,5)
      OPEN 1,NAME,LEN=4
      TYPE "EIGENVALUES PLUS AVERAGE # OF FRAMES "
      DO 115 K=1,NP
      WRITE(1) EIG(K)
      TYPE EIG(K)
115   CONTINUE
      AV=MEAN
      WRITE(1) AV
      TYPE AV
      CLOSE 1
      GO TO 105
999   IF(LAM.NE.0) GO TO 105
      TYPE "ENTER FILE WITH REFERENCE EIGENVECTORS "
      READ(11,5) (NAME(I),I=1,5)
      NPN=NP*NP
      OPEN 4,NAME
      READ(4) (R(L),L=1,NPN)
      WRITE(12,6) (R(L),L=1,NPN)
      CLOSE 4
      DO 212 I=1,NP
      DO 212 K=1,NP
      L=(I-1)*NP+K
      E(K,I)=R(L)
212   CONTINUE
C
C       CALCULATION OF ORTHOGONAL PARAMETERS
C
105   DO 25 J=1,NF
      DO 25 I=1,NP
      SUM1=0.0
      DO 95 K=1,NP
      SUM1= SUM1+E(K,I)*B(K,J)
 95   CONTINUE
      ORT(I,J)=SUM1
 25   CONTINUE
```

```
C
C       CALCULATION THE AVERAGE VALUE OF THE I-TH ORTHONAL PARAMETER
C       ACROSS THE UTTERANCE
C
        DO 75, I=1, NP
        AVER=0.0
        DO 85 J=1, NF
   85   AVER=AVER+ORT(I, J)
        AVORT(I)=AVER
   75   CONTINUE
C
C       CALCULATION THE AVERAGE VALUE OF I-TH ORTHOGONAL PARAMETER
C       FOR A GIVEN SPEAKER
C
        DO 106 I=1, NP
        AVERQ(I)=AVERQ(I)+AVORT(I)
  106   CONTINUE
        IF(III.EQ.3) GO TO 333
        ACCEPT "WISH TO CONTINUE PROCESSING ? YES-1, NO-0", IND
        IF(IND.EQ.1) GO TO 55
    7   FORMAT("0", 6E15.7)
  333   DO 110 K=1, NP
        AVERQ(K)=AVERQ(K)/MSUM
        TYPE AVERQ(K)
  110   CONTINUE
C
        ACCEPT "WISH TO STORE REFERENCE FILE WITH ORTH. PAR.: ?YES-1, NO-0", MARK
        IF(MARK.NE.1) GO TO 100
        TYPE "ENTER OUTPUT REFERENCE FILENAME"
        READ(11, 5) (NAME(I), I=1, 5)
        OPEN 2, NAME, LEN=4
        DO 116 K=1, NP
        WRITE(2) AVERQ(K)
  116   CONTINUE
        CLOSE 2
        LAM=LAM+1
        ACCEPT "WISH TO REPEAT THE WHOLE OPERATION -? YES-1, NO-0", LIL
        IF(LIL.EQ.1) GO TO 220
  100   STOP
        END
```

```
      SUBROUTINE EIGEN(A,R,N,NA,NR,MV)
C
C        THIS SUBROUTIN  COMPUTES THE EIGEN VALUES AND EIGEN VECTORS OF MATRIX A
C
C        DESCRIPTION OF PARAMETERS
C        A - ORIGINAL MATRIX(SYMMETRIC), DESTROYED IN COMPUTATION.
C            RESULTANT EIGEN VALUES ARE DEVELOPED IN DIAGONAL OF
C            MATRIX A IN DESCENDING ORDER.
C        R - RESULTANT MATRIX OF EIGENVECTORS (STORED COLUMNWISE,
C            IN SAME SEQUENCE AS EIGEN VALUES)
C        MV - INPUT CODE
C                  0   COMPUTE EIGEN VALUES AND EIGEN VECTORS
C                  1   COMPUTE EIGENVALUES ONLY, R NEED NOT BE DIMENSIONED
C                      IN THE MAIN PROGRAM BUT MUST APPEAR IN THE CALLING SEQNCE.
C        NA=N*(N+1)/2      SHOULD BE DEFINED IN MAIN PROGRAM
C        NR=N*N            SHOULD BE DEFINED IN MAIN PROGRAM
C
      DIMENSION A(NA),R(NR)
    5 RANGE=1.0E-6
      IF(MV-1) 10,25,10
   10 IQ=-N
      DO 20 J=1,N
      IQ=IQ+N
      DO 20 I=1,N
      IJ=IQ+I
      R(IJ)=0.0
      IF(I-J) 20,15,20
   15 R(IJ)=1.0
   20 CONTINUE
C
C        COMPUTE INITIAL AND FINAL NORMS (ANORM AND ANORMX)
C
   25 ANORM=0.0
      DO 35 I=1,N
      DO 35 J=I,N
      IF(I-J) 30,35,30
   30 IA=I+(J*J-J)/2
      ANORM=ANORM+A(IA)*A(IA)
   35 CONTINUE
      IF(ANORM) 165,165,40
   40 ANORM=1.414*SQRT(ANORM)
      ANRMX=ANORM*RANGE/FLOAT(N)
C
C        INITIALIZE INDICATORS AND COMPUTE THRESHOLD, THR
C
      IND=0
      THR=ANORM
   45 THR=THR/FLOAT(N)
   50 L=1
   55 M=L+1
C
```

```
C
C          COMPUTE SIN AND COS
C
   60 MQ=(M*M-M)/2
      LQ=(L*L-L)/2
      LM=L+MQ
   62 IF( ABS(A(LM))-THR) 130,65,65
   65 IND=1
      LL=L+LQ
      MM=M+MQ
      X=0.5*(A(LL)-A(MM))
   68 Y=-A(LM)/ SQRT(A(LM)*A(LM)+X*X)
      IF(X) 70,75,75
   70 Y=-Y
   75 SINX=Y/ SQRT(2.0*(1.0+( SQRT(1.0-Y*Y))))
      SINX2=SINX*SINX
   78 COSX= SQRT(1.0-SINX2)
      COSX2=COSX*COSX
      SINCS =SINX*COSX
C
C          ROTATE L AND M COLUMNS
C
      ILQ=N*(L-1)
      IMQ=N*(M-1)
      DO 125 I=1,N
      IQ=(I*I-I)/2
      IF(I-L) 80,115,80
   80 IF(I-M) 85,115,90
   85 IM=I+MQ
      GO TO 95
   90 IM=M+IQ
   95 IF(I-L) 100,105,105
  100 IL=I+LQ
      GO TO 110
  105 IL=L+IQ
  110 X=A(IL)*COSX-A(IM)*SINX
      A(IM)=A(IL)*SINX+A(IM)*COSX
      A(IL)=X
  115 IF(MV-1) 120,125,120
  120 ILR=ILQ+I
      IMR=IMQ+I
      X=R(ILR)*COSX-R(IMR)*SINX
      R(IMR)=R(ILR)*SINX+R(IMR)*COSX
      R(ILR)=X
  125 CONTINUE
      X=2.0*A(LM)*SINCS
      Y=A(LL)*COSX2+A(MM)*SINX2-X
      X=A(LL)*SINX2+A(MM)*COSX2+X
      A(LM)=(A(LL)-A(MM))*SINCS+A(LM)*(COSX2-SINX2)
      A(LL)=Y
      A(MM)=X
```

```
C
C         TESTS FOR COMPLETION
C
C         TEST FOR M = LAST COLUMN
C
  130 IF(M-N) 135,140,135
  135 M=M+1
      GO TO 60
C
C         TEST FOR L = SECOND FROM LAST COLUMN
C
  140 IF(L-(N-1)) 145,150,145
  145 L=L+1
      GO TO 55
  150 IF(IND-1) 160,155,160
  155 IND=0
      GO TO 50
C
C         COMPARE THRESHOLD WITH FINAL NORM
C
  160 IF(THR-ANRMX) 165,165,45
C
C         SORT EIGENVALUES AND EIGENVECTORS
C
  165 IQ=-N
      DO 185 I=1,N
      IQ=IQ+N
      LL=I+(I*I-I)/2
      JQ=N*(I-2)
      DO 185 J=I,N
      JQ=JQ+N
      MM=J+(J*J-J)/2
      IF(A(LL)-A(MM)) 170,185,185
  170 X=A(LL)
      A(LL)=A(MM)
      A(MM)=X
      IF(MV-1) 175,185,175
  175 DO 180 K=1,N
      ILR=IQ+K
      IMR=JQ+K
      X=R(ILR)
      R(ILR)=R(IMR)
  180 R(IMR)=X
  185 CONTINUE
      RETURN
      END
```

```
C              "ORTSUM"
C         PROGRAM TO COMPUTE REFERENCE ORTHOGONAL PARAMETERS
C
      DIMENSION X(20),Y(20),NAME(5)
 1    ACCEPT "THE ORDER OF LPC MODEL=",NP
      D=0.
      DO 7 I=1,NP
      Y(I)=0.0
 7    CONTINUE
 10   TYPE "ENTER INPUT FILENAME WITH ORT. PAR."
      READ(11,5) (NAME(I),I=1,5)
 5    FORMAT(5A2)
      OPEN 0, NAME,LEN=4*NP
      READ(0) (X(K),K=1,NP)
      CLOSE 0
      DO 8 K=1,NP
      Y(K)=Y(K)+X(K)
 8    TYPE Y(K)
      D=D+1.
      ACCEPT "WISH TO READ ANOTHER FILE ? YES-1,NO-0 ",JA
      IF(JA.EQ.1) GO TO 10
      DO 15 I=1,NP
      Y(I)=Y(I)/D
      TYPE Y(I)
 15   CONTINUE
 20   ACCEPT "WISH TO STORE REFERENCE ORT. PAR.? YES-1,NO-0 ",MARK
      IF(MARK.NE.1) GO TO 30
      TYPE "ENTER OUTPUT REFERENCE FILENAME"
      READ(11,5) (NAME(I),I=1,5)
      OPEN 1,NAME,LEN=4
      DO 25 I=1,NP
      WRITE(1) Y(I)
 25   CONTINUE
      CLOSE 1
 30   STOP
      END
```

```
      " DIST "
C
C     PROGRAM COMPUTES DISTANCE USING DIFFERENT NUMBER OF ORHOGONAL PARAMETERS
C
      DIMENSION NAME(5),X(16),Y(17),Z(16)
      ACCEPT "THE ORDER OF LPC MODEL=",NP
      ACCEPT "WANT TO PRINT OUT RESULTS ? YES-1,NO-0",IW
      NP1=NP+1
      DO 1 I=1,NP
1     X(I)=0.0
      TYPE "ENTER INPUT REFERENCE FILE WITH EIGENVALUES"
      READ(11,5) (NAME(I),I=1,5)
5     FORMAT(5A2)
      OPEN 0,NAME,LEN=4*(NP1)
      READ(0) (Y(I),I=1,NP1)
      IF(IW.NE.1) GO TO 22
      WRITE(12,24) (Y(I),I=1,NP1)
22    CLOSE 0
      DO 11 I=1,NP1
      TYPE Y(I)
11    CONTINUE
      TYPE "ENTER INPUT REFERENCE FILE WITH ORTHOGONAL PARAMETERS"
      READ(11,5) (NAME(I),I=1,5)
      OPEN 1,NAME,LEN=4*NP
      READ(1) (X(I),I=1,NP)
      IF(IW.NE.1) GO TO 26
      WRITE(12,24) (X(I),I=1,NP)
26    CLOSE 1
      DO 21 I=1,NP
      TYPE X(I)
21    CONTINUE
      TYPE "COMPUTE DISTANCE INCLUDING ONLY ORT. PAR. IN THE RANGE N<K<NN"
      ACCEPT " N=",N
      ACCEPT " NN=",NN
45    D=0.0
      TYPE "ENTER INPUT FILE WITH ORTHOGONAL PARAMETERS OF
     * AN UNKNOWN SPEAKER"
      READ(11,5) (NAME(I),I=1,5)
      OPEN 2,NAME,LEN=4*NP
      READ(2) (Z(I),I=1,NP)
      IF(IW.NE.1) GO TO 28
      WRITE(12,24) (Z(I),I=1,NP)
28    CLOSE 2
      DO 31 I=1,NP
      TYPE Z(I)
31    CONTINUE
      DO 16 L=N,NN
      D=0.0
      DO 15 K=L,NN
      D=D+((X(K)-Z(K))**2)/Y(K)
15     CONTINUE
       AV=Y(NP1)
      D=D*AV
      TYPE "DISTANCE IS EQUAL ",D
      IF(IW.NE.1) GO TO 16
      WRITE(12,25) D
16    CONTINUE
      ACCEPT "WISH TO CONTINUE PROCESSING ? YES-1,NO-0",IND
      IF(IND.EQ.1) GO TO 45
25    FORMAT(5X,"DISTANCE =",F15.5)
24    FORMAT("0",3X,8E15.7)
      STOP
      END
```

```
      "NOISE"
C        PROGRAM - GENERATES PSEUDO-RANDOM GAUSSIAN NOISE
C                - ADDS IT TO CLEAN SPEECH SIGNAL AND CREATES NOISY SPEECH
C                  WITH DESIRED SIGNAL-TO-NOISE RATIO WITHIN 0-30 DB
C                - COMPUTES LPC PARAMETERS FROM HIGH QUALITY AND NOISY SPEECH
C     SUBROUTINES REQUIRED:GAUSS,AKN1,AKN2
      DIMENSION A(16),R(16),X(300),Y(300),YY(300)
      DIMENSION XX(300),AA(16),AAA(16)
      REAL SQRT,FLOAT
      INTEGER NAME(5), IS(300),IN(300),ISN(300),IZ(300)
      ACCEPT "# OF LPC=",NPP
      ACCEPT "LENGTH OF FILE ",ND
      ACCEPT "# OF FRAMES ",NS
      ACCEPT "S/N-RATIO IN DB =",NX
      XN=0.1*FLOAT(NX)
      RATIO=10**XN
      TYPE RATIO
      NFT=ND/NS
      TYPE "ENTER INPUT FILENAME "
      READ(11,5) (NAME(I),I=1,5)
 5    FORMAT(5A2)
      OPEN 0,NAME
      ACCEPT "WISH TO STORE LPC'S OF HI-FI SPEECH ? YES-1,NO-0 ",MARK
      IF (MARK.NE.1) GO TO 105
      TYPE "ENTER OUTPUT FILE WITH LPC'S OF HI-FI SPEECH"
      READ(11,5) (NAME(I),I=1,5)
      OPEN 1,NAME,LEN=4*(NPP+1)
 105  TYPE "ENTER OUTPUT FILE WITH LPC'S OF NOISY SPEECH"
      READ(11,5) (NAME(I),I=1,5)
      OPEN 2,NAME,LEN=4*(NPP+1)
      TYPE "ENTER OUTPUT FILE WITH LPC'S OF SPEECH WITH ATTENUATED NOISE"
      READ(11,5) (NAME(I),I=1,5)
      OPEN 3,NAME,LEN=4*(NPP+1)
      DO 11 KK=1,NS
      IF(KK.EQ.NS) NFT=ND-(KK-1)*NFT
      READ(0) (IS(J),J=1,NFT)
 30   FORMAT(16I5)
      SUM=0.
      AVPOW=0.0
      DO 1 I=1,NFT
      IS(I)=10*IS(I)
      XX(I)=I
      X(I)=IS(I)
 1    SUM=SUM+IS(I)
      SUM=SUM/NFT
      DO 14 I=1,NFT
      IS(I)=IS(I)-SUM
      SS=IS(I)
 14   AVPOW=AVPOW+SS**2
      AVPOW=AVPOW/NFT
      VAR=AVPOW/RATIO
      STDEV=SQRT(VAR)
      TYPE "VARIANCE OF NOISE=",VAR
      IX=15379
      AM=0.0
      DO 8 J=1,NFT
      CALL GAUSS(IX,STDEV,AM,V)
      IN(J)=V
 8    CONTINUE
```

```
      DO 2 I=1,NFT
      ISN(I)=IS(I)+IN(I)
      Y(I)=ISN(I)
    2 CONTINUE
      WRITE(6,10)
   10 FORMAT(/)
C     LPC ANALYSIS
C     EXAMPLE:    S/N RATIO= 30 DB          SCALING FACTOR=AVPOW/1000
      DO 20 J=1,NFT
      Y(J)=ISN(J)
   20 X(J)=IS(J)
      CALL AKN1(X,A,R,NPP,NFT,AL)
      DO 85 K=1,NPP
   85 AA(K)=A(K)
      CALL AKN1(Y,A,R,NPP,NFT,AL)
      DO 90 K=1,NPP
   90 AAA(K)=A(K)
      WRITE(12,25) (AA(J),J=1,NPP)
      WRITE(12,25) (AAA(J),J=1,NPP)
      ZZ=NS
      IF (MARK.NE.1) GO TO 120
      WRITE(1) ZZ,(AA(J),J=1,NPP)
  120 WRITE(2) ZZ,(AAA(J),J=1,NPP)
      CALL AKN2(Y,A,R,NPP,NFT,AL,VAR)
      WRITE(3) ZZ,(A(J),J=1,NPP)
      WRITE(12,25) (A(J),J=1,NPP)
   11 CONTINUE
      CLOSE 8
      IF(MARK.NE.1) GO TO 110
      CLOSE 1
  110 CLOSE 2
      CLOSE 3
   25 FORMAT('0',3X,8E15.7)
      STOP
      END
```

```
          "SPECTRUM"
C
C       PROGRAM COMPUTES SPEECH SIGNAL SPECTRUM ESTIMATE
C       AND LPC MODEL SPECTRUM ESTIMATE USING FFT-TECHNIQUE
C       HAMMING WINDOW IS USED OR NOT
C       X-SPEECH SAMPLES
C       SUBROUTINE REQUIRED:FFT,AKN1,FLTEKS
C
        REAL X(256),Y(256),Z(256),XLM(256),R(16),A(16),CK(20),XJ(256)
        REAL ZA(256),COS,ALOG10,ABS,ZZA(256)
        INTEGER NAME(5),IX(16716)
        LCOUNT=0.0
  1     OPEN 0,"$TTO1"
        ACCEPT "# OF INPUT SAMPLES=",NS
        ACCEPT "ORDER OF LPC MODEL ",NPP
        ACCEPT "# OF FRAMES ",NF
        TYPE " 256-POINTS FFT "
        ACCEPT"WISH TO ANALYSE INPUT DATA OR MODEL [1/A(2)]
     X  IF DATA -1, IF MODEL -0 ",LINK
        ACCEPT "WISH TO ANALYSE ONLY ONE SEGMENT OF SPEECH ? YES-1,NO-0",IND
        IF(IND.NE.1) GO TO 35
        ACCEPT "BEGINNING OF THE ANALYSED SEGMENT ",NNN
  35    L=8
        NN=2**L
        ACCEPT" WISH TO USE HAMMING WINDOW ? YES-1,NO-0",MARK
        NSF=NS/(NF-1)
        SUM=0.0
        GAIN=0.0
        DO 4 I=1,NN
  4     ZA(I)=0.0
        TYPE "ENTER INPUT FILE "
        READ(11,2)(NAME(I),I=1,5)
  2     FORMAT(5A2)
        OPEN 1,NAME
        READ(1) (IX(J),J=1,NS)
        NSFF=NSF
        NF1=NF
        IF(IND.EQ.1) NF1=1
        DO 10 K=1,NF1
        IF(IND.EQ.1) GO TO 60
        IF(K.EQ.NF1)NSF=NS-(K-1)*NSF
        DO 7 J=1,NSF
        Y(J)=0.0
        JJ=(K-1)*NSFF+J
        X(J)=IX(JJ)
  7     CONTINUE
        GO TO 65
  60    DO 17 J=1,NN
        Y(J)=0.0
  17    X(J)=IX(NNN+J-1)
        NSF=NN
  65    IF(MARK.NE.1) GO TO 70
        PIN=6.28318/NSF
        DO 75 J=1,NSF
  75    X(J)=(X(J))*(0.54+0.46*COS(J*PIN))
  70    CALL AKN1(X,A,R,NPP,NSF,AL)
        TYPE "AKN1 IS O'KEY"
```

```
      NIP=NPP+1
      DO 44 II=1,NIP
   44 GAIN=GAIN+A(II)*R(II)
      IF(LINK.EQ.1) GO TO 45
      NSF1=NPP+2
      X(1)=1.
      Y(1)=0.0
      DO 15 J=2,NN
      X(J)=0.0
      Y(J)=0.0
      IF(J.LT.NSF1) X(J)=A(J-1)
   15 CONTINUE
   45 CALL FFT(X,Y,L)
      TYPE "FFT IS O'KEY"
      DO 20 J=1,NN
      Z(J)=(X(J))**2+(Y(J))**2
      ZA(J)=Z(J)+ZA(J)
   20 CONTINUE
      SUM=SUM+AL
   10 CONTINUE
      AL=SUM/NF1
      AL1=20.*ALOG10(AL)
      TYPE "GAIN=",AL,GAIN
      NN=NN/2+1
      XMAX=0.0
      DO 25 J=1,NN
      ZA(J)=ZA(J)/NF1
      XLM(J)=10.*ALOG10(ZA(J))
      IF(LINK.NE.1) XLM(J)=-XLM(J)
      IF(XLM(J).GT.XMAX)XMAX=XLM(J)
      XJ(J)=J*31.25
   25 CONTINUE
      TYPE "MAX. LOG MAGNITUDE=",XMAX
      DO 50 I=1,NN
      XLM(I)=XLM(I)-XMAX
      WRITE(12,30) I,XJ(I),XLM(I)
   50 CONTINUE
   30 FORMAT(" ",I5,F10.2,1X,"HZ",2X,E18.10)
      CALL PLTEKS(XJ,XLM,NN)
      CLOSE 1
      CLOSE 0
      SUMA1=0.0
```

```
      DO 95 I=1,NN
      ZA(I)=ABS(ZA(I))
95    SUMA1=SUMA1+(ZA(I))**2
      IF(LCOUNT.EQ.1) GO TO 100
      HIFI=SUMA1
      DO 55 I=1,NN
55    ZZA(I)=ZA(I)
      LCOUNT=LCOUNT+1
      GO TO 1
100   SUMA2=0.0
      DO 85 I=1,NN
85    SUMA2=SUMA2+(ABS(ZA(I)-ZZA(I)))**2
      TYPE "HIFI=",HIFI
      TYPE "SUMA2=",SUMA2
      STON=10.*ALOG10(HIFI/SUMA2)
      TYPE "S/N-RATIO IN DB =",STON
      STOP
      END
```

```
C     RADIX-2 FAST FOURIER TRANSFORM (FFT)
C
      SUBROUTINE FFT(X,Y,L)
      DIMENSION X(256),Y(256)
      NP=2**L
      LMX=NP
      SCL=6.283185303/NP
      DO 20 LO=1,L
      LIX=LMX
      LMX=LMX/2
      ARG=0.
      DO 10 LM=1,LMX
      C=COS(ARG)
      S=SIN(ARG)
      ARG=ARG+SCL
      DO 10 LI=LIX,NP,LIX
      J1=LI-LIX+LM
      J2=J1+LMX
      T1=X(J1)-X(J2)
      T2=Y(J1)-Y(J2)
      X(J1)=X(J1)+X(J2)
      Y(J1)=Y(J1)+Y(J2)
      X(J2)=C*T1+S*T2
 10   Y(J2)=C*T2-S*T1
 20   SCL=2.*SCL
C
C     BIT REVERSAL
C
      J=1
      NV2=NP/2
      NPM1=NP-1
      DO 50 I=1,NPM1
      IF(I.GE.J) GO TO 30
      T1=X(J)
      T2=Y(J)
      X(J)=X(I)
      Y(J)=Y(I)
      X(I)=T1
      Y(I)=T2
 30   K=NV2
 40   IF (K.GE.J) GO TO 50
      J=J-K
      K=K/2
      GO TO 40
 50   J=J+K
      RETURN
      END
```

```
C       PITCHEXT1
C
C       PROGRAM COMPUTES PITCH PERIODS
C
        DIMENSION NPBUF(100),IBUF1(2000),P(110),NAME(5),PBUF(100),PITCH(3)
        DIMENSION SPCH(400)
        INTEGER FRAME
        NBUF=600
        NY=400
        ACCEPT '# OF INPUT SAMPLES!',NNN
        ACCEPT "DOWN SAMPLING FACTOR=",ID
        NKN=200/ID
        TYPE "ENTER INPUT FILENAME"
        READ(11,1) (NAME(I),I=1,5)
    1   FORMAT(5A2)
        KK=NBUF
        FRAME=0
        OPEN 0,NAME
        READ(0)  (IBUF1(I),I=1,NBUF)
        TYPE "PITCH FILENAME IS"
        READ(11,1) (NAME(I),I=1,5)
        OPEN 1,NAME
        ACCEPT "WISH TO STORE DOWN SAMPLED SPEECH-?,YES-1,NO-0",MARK
        IF(MARK.NE.1) GO TO 4
        TYPE "FILENAME WITH DOWN SAMPLED SPEECH"
        READ(11,1) (NAME(I),I=1,5)
        OPEN 2,NAME
    4   K=0
        FRAME=FRAME+1
        IF(FRAME.NE.1) GO TO 7
        IF(FRAME.EQ.1) L=0
        DO2J=1,3
    2   PITCH(J)=0
    7   DO6J=1,400
        SPCH(J)=IBUF1(J+K)/500.
    6   CONTINUE
        CALL STEP1(SPCH,PBUF,ID)
        DO 11 J=1,NKN
   11   NPBUF(J)=PBUF(J)
        IF(MARK.NE.1) GO TO 8
        WRITE(2) (NPBUF(J),J=1,NKN)
    8   CALL STEP2(PBUF,PITCH)
        K=K+200
        L=L+1
        IF(L.GE.3)P(L-2)=PITCH(3)
        IF(K+400.LE.NBUF)GOTO7
        IF(NNN-KK.LT.400) NY=NNN-KK
        IF(KK.GE.NNN) GO TO 5
        DO 10 I=1,200
   10   IBUF1(I)=IBUF1(I+400)
        READ(0) (IBUF1(I+200),I=1,NY)
        IF(NY.EQ.400) GO TO 13
        NY=NY+1
```

```
      DO 12 I=NY,400
   12 IBUF1(I+200)=0
   13 KK=KK+400
      GO TO 4
    5 CLOSE 0
      P(L-1)=PITCH(2)
      P(L)=PITCH(1)
      WRITE(1) (P(J),J=1,L)
      IF(NY.GE.200) P(L+1)=0.
      WRITE(1) P(L+1)
      CLOSE 1
      IF(MARK.NE.1) GO TO 9
      CLOSE 2
      L=L+1
    9 WRITE(12,3) (P(J),J=1,L)
    3 FORMAT('0',F10.6)
      STOP
      END
```

```
C               "CANCEL"
C
      INTEGER NAME(5),IY(200),ISR(200)
      DIMENSION P(100),A(16),YIY(11500),E(200),SR(200)
      ACCEPT" ORDER OF LPC MODEL",NP
      ACCEPT "LENGTH OF INPUT FILE ",NN
      ACCEPT "LENGTH OF PITCH BLOCKS ",NS
      ID=5
      LB=NN/NS
      NRES=NN-NS*LB
      LB=LB+1
      NPP1=NP+1
      BETA=2*1.E-17
      NSUMA=0
      ICOUNT=1
      K1=1
      K2=NS
      A(1)=0.077
      DO 1 I=2,NPP1
1     A(I)=0.077
      TYPE "ENTER INPUT FILENAME"
      READ(11,5) (NAME(I),I=1,5)
5     FORMAT(5A2)
      OPEN 0,NAME.
      TYPE "ENTER FILE WITH PITCH "
      READ(11,5) (NAME(I),I=1,5)
      OPEN 1,NAME,LEN=4*LB
      TYPE "ENTER OUTPUT FILENAME"
      READ(11,5) (NAME(I),I=1,5)
      OPEN 2,NAME,LEN=2*NS
      READ(1) (P(I),I=1,LB)
      P(1)=0.
10    TYPE "CYCLE # ",ICOUNT
      IF(ICOUNT.EQ.LB)NS=NRES
      READ(0) (IY(J),J=1,NS)
C     WRITE(12,600) (IY(J),J=1,NS)
      NT=(P(ICOUNT)-1)*ID
      IF(NT.LE.0)NT=0
      TYPE "PITCH=",NT
      IF(ICOUNT.GT.1)K1=NS+1
      IF(ICOUNT.GT.1)K2=2*NS
      DO 200 K=K1,K2
      SUM1=0.0
      YIY(K)=IY(K-K1+1)
      DO 500 N=1,NPP1
      IF(ICOUNT.GT.1) GO TO 500
      IF((K.LE.(N+NT)).AND.(ICOUNT.EQ.1)) SR(K)=YIY(K)
      IF((K.LE.(N+NT)).AND.(ICOUNT.EQ.1)) GO TO 2
500   SUM1=SUM1+(A(N))*YIY(K-N-NT)
```

```
      SR(K-K1+1)=SUM1
      IF(NT.LE.0) SR(K-K1+1)=0.1*YIY(K)
      IF(NT.LE.0) GO TO 200
    2 E(K-K1+1)=YIY(K)-SR(K-K1+1)
      IF(K.GT.NT) CONST=BETA*(E(K-K1+1))*YIY(K-NT)
      IF(K.LE.NT) CONST=0.0
      DO 300 N=1,NPP1
  300 A(N)=A(N)+CONST
C     WRITE(12,400) (A(N),N=1,NPP1)
  200 CONTINUE
      DO 250 J=1,NS
  250 ISR(J)=SR(J)
C     WRITE(12,600) (ISR(J),J=1,NS)
      WRITE(2) (ISR(J),J=1,NS)
C     WRITE(12,700)
  700 FORMAT(//)
      IF(ICOUNT.GE.LB) GO TO 900
      ICOUNT=ICOUNT+1
      IF(ICOUNT.EQ.2)GO TO 10
      DO 800 J=1,NS
  800 YIY(J)=YIY(J+NS)
      GO TO 10
  900 CLOSE 0
      CLOSE 1
      CLOSE 2
  400 FORMAT("0",3X,8E15.7)
  600 FORMAT(" ",2X,20I6)
      STOP
      END
```

BIBLIOGRAPHY

1. J.D. Markel, A.H. Gray, Jr., "Linear Prediction of Speech," Springer-Verlag Berlin Heidelberg: New York, 1976.

2. B.S. Atal & S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., 50, pp. 637-655, 1971.

3. M.R. Sambur, "An Efficient Linear-Prediction Vocoder", Bell System Tech. J., Vol. 54, No. 10, Dec. 1975.

4. R.C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio & Electroacoustics, Vol. AU-21, No. 2, pp. 80-88, Apr. 1973.

5. G.R. Doddington, "A computer method of speaker verification," PhD dissertation, Dept. Elect. Eng., Univ. Wisconsin, Madison, 1970.

6. Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-23, pp. 67-72, 1975.

7. L.L. Pfeifer, "Inverse Filter for Speaker Identification. RADC-TR-74-214, Final Report, Speech Communications Research Laboratory, Santa Barbara, Ca., 1974.

8. B.S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification. J. Acoust. Soc. Am. 55, pp. 1304-1312, 1974.

9. H. Cramer, "Mathematical Methods in Statistics," Princeton Univ. Press, 1951.

10. M.R. Sambur, "Speaker Recognition Using Orthogonal Linear Prediction," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 4, Aug. 1976.

11. A.E. Rosenberg, "Evaluation of an Automatic Speaker-Verification System Over Telephone Lines," Bell System Tech. J., Vol. 55, No. 6, pp. 723-743, 1976.

12. G.R. Doddington, "A method of speaker verification," J. Acoust. Soc. Amer. Vol. 49, 1971.

13. M.R. Sambur and N.S. Jayaut "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise", IEEE, Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 6, Dec. 1976.

14. M. Shridhar and M. Baraniecki, "Accuracy of Speaker Verification via Orthogonal Prameters for Noisy Speech," 1979 IEEE Proc. Int. Conf., Acoust., Speech, Signal Processing, pp. 785-788, Apr. 2-4, 1979.

15. Steven M. Kay, "The Effects of Noise on the Auto-regressive Spectral Estimator," IEEE, Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, No. 5, Oct. 1979.

16. Widrow et at., "Adaptive Noise Cancelling: Principles and Applications" Proc. IEEE, Vol. 63, pp. 1692-1716, Dec. 1975.

17. M.R. Sambur, "Adaptive Noise Cancelling for Speech Signals," Trans. Acoust., Speech and Signal Processing, Vol. ASSP-26, No. 5, Oct. 1978.

18. M. Baraniecki and M. Shridhar, "A Speaker Verification Algorithm for Speech Utterances Corrupted by Noise with Unknown Statistics," 1980 IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing, April 9-11, 1980.

19. S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans.

20. B. Beck et al, "An Assessment of the Technology of Automatic Speech Recognition for Military Applications," IEEE Trans., ASSP. 25, No. 4, Aug. 77.

21. A.E. Rosenberg and M.R. Sambur, "New Techniques for Automatic Speaker Verification," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-23, No. 2, April, 1975.

22. Lo-Soun, K.-P. Li, and K.S. Fu, "Identification of Speakers by Use of Nasal Coarticulation," J. Acoust. Soc. Am., Vol. 56, Dec. 1974.

23. Bolt, R.H., et al., "Speaker Identification by Speech Spectrograms: A Scientist's View of its Reliability for Legal Purposes," J. Acoust. Soc. Amer., Vol. 47, pp. 597-612, 1970.

24. Stevens, K.N., et al., "Speaker Authentication and Identification: A comparison of Spectrographic and Auditory Presentations of Speech Material," J.Acoust. Soc. Amer., Vol. 44, pp. 1596-1603, 1968.

25. S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," J. Acoust. Soc. Amer., Vol. 35, pp. 354-358, Mar. 1963.

26. P.D. Bricker et al., "Statistical techniques for talker identification," Bell Syst. Techn. J., Vol. 50, pp. 1427-1454, Apr. 1971.

27. B.S. Atal, "Automatic Speaker Recognition based on pitch contours," J. Acoust. Soc. Amer., Vol. 52, pp. 1687-1697, Dec. 1972.

28. J.J. Wolf, "Efficient acoustic parameters for speaker recognition," J. Acoust. Soc. Amer., Vol. 51, pt. 2, pp. 2044-2055, 1971.

29. Li, K-P, et al., "Experimental Studies in Speaker Verification Using an Adaptive System," J. Acoust. Soc. Amer., Vol. 40, pp. 966-978, 1966.

30. Das, S.K. and Mohn, W.S., "A Scheme for Speech Processing in Automatic Speaker Verification," IEEE Trans. Audio Electroacoust., Vol. AU-19, pp. 32-43, 1971.

31. R.C. Lummis and A.E. Rosenberg, "Test of an automatic speaker verification method with intensively trained mimics," J. Acoust. Soc. Amer., Vol. 51, p. 131(A), 1972.

32. G.R. Doddington, "Personal Identity Verification Using Voice," Proc. ELECTRO-76, pp. 22-4, 1-5 May 11-14, 1976.

33. R.W. Schaefer and L.R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech, J. Acoust. Soc. Amer., Vol. 47, pp. 634-648, Feb. 1970.

34. S.S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," IEEE. Trans. Acoust. Speech, Signal Processing, Vol. ASSP-22, pp. 135-141, Apr. 1974.

35. K.P. Li, G.W. Hughes, and A.S. House, "Correlation characteristics and dimensionality of speech spectra," J. Acoust. Soc. Amer., Vol. 46, pt. 2, pp. 1019-1025, Oct. 1969.

36. J.D. Markel and S.B. Davis, "Text-independent Speaker Identification from a Large Linguistically Unconstrained Time-Spaced Data Base," 1978 IEEE. Proc. Int. Conf. ASSP, May, 1978.

37. C.A. McGonegal, L. Rabiner, and A.E. Rosenberg, "A subjective evaluation of pitch detection methods using LPC synthesized speech," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, pp. 221-229, June 1977.

38. S. Maitra and C.R. Davis, "A Speech Digitzer at 2400 Bits/s", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, No. 6, pp. 729-733, Dec. 1979.

39. L.R. Rabiner and B. Gold "Theory and Application of Digital Signal Processing," Prentice-Hall, Inc., 1976.

40. L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Inc., Englewood Cliffs, 1978.

41. M. Vidalon, "Effectiveness of Low Frequency Parameters for Automatic Speaker Verification," Ph.D. Dissertation, Dept. Elect. Eng., University of Windsor, Windsor, Canada.

VITA AUCTORIS

1945	Born in Kielce, Poland
1963	Graduated from S. Zeromski High School, Kielce, Poland
1970	Graduated from Technical University of Warsaw, Electronics Department, Warsaw, Poland, with Ing.Dipl. and M.Sc. degree in Electronics
1970-76	Employed in the Institute of Telecommunications, Warsaw, Poland; engaged in the field of pulse code modulation, delta modulation techniques and digital transmission
1980	Candidate for Ph.D. degree, Electrical Engineering Department, University of Windsor, Windsor, Ontario, Canada